Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/JP05/000505

International filing date: 17 January 2005 (17.01.2005)

Document type: Certified copy of priority document

Document details: Country/Office: JP
Number: 2004-018715
Filing date: 27 January 2004 (27.01.2004)

Date of receipt at the International Bureau: 10 March 2005 (10.03.2005)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)

PCT/JP2005/000505

日 本 国 特 許 庁
JAPAN PATENT OFFICE

20.01.2005

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

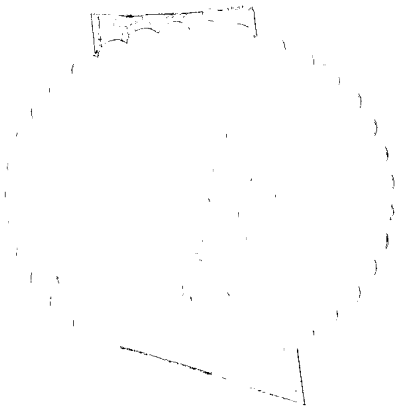
This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: ２００４年　１月２７日

出 願 番 号
Application Number: 特願２００４－０１８７１５

[ST. 10/C]: ［ＪＰ２００４－０１８７１５］

出 願 人
Applicant(s): 松下電器産業株式会社



２００５年　２月２５日

特許庁長官
Commissioner,
Japan Patent Office

小 川 洋

出証番号　出証特２００５－３０１５０９５

【書類名】　特許願
【整理番号】　2033850244
【あて先】　特許庁長官殿
【国際特許分類】　G10L 13/00
【発明者】
　【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内
　【氏名】　齋藤　夏樹
【発明者】
　【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内
　【氏名】　釜井　孝浩
【発明者】
　【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内
　【氏名】　加藤　弓子
【特許出願人】
　【識別番号】　000005821
　【氏名又は名称】　松下電器産業株式会社
【代理人】
　【識別番号】　100109210
　【弁理士】
　【氏名又は名称】　新居　広守
【手数料の表示】
　【予納台帳番号】　049515
　【納付金額】　21,000円
【提出物件の目録】
　【物件名】　特許請求の範囲　1
　【物件名】　明細書　1
　【物件名】　図面　1
　【物件名】　要約書　1
　【包括委任状番号】　0213583

【書類名】特許請求の範囲
【請求項１】
第１の声質に属する複数の音声素片に関する第１の音声素片情報、及び前記第１の声質と異なる第２の声質に属する複数の音声素片に関する第２の音声素片情報を予め記憶している記憶手段と、
テキストデータを取得するとともに、前記記憶手段の第１の音声素片情報から、前記テキストデータに含まれる文字に対応した前記第１の声質の合成音声を示す第１の合成音声情報を生成し、前記記憶手段の第２の音声素片情報から、前記テキストデータに含まれる文字に対応した前記第２の声質の合成音声を示す第２の合成音声情報を生成する音声情報生成手段と、
前記音声情報生成手段により生成された前記第１及び第２の合成音声情報から、前記テキストデータに含まれる文字に対応した、前記第１及び第２の声質の中間的な声質の合成音声を示す中間合成音声情報を生成するモーフィング手段と、
前記モーフィング手段によって生成された前記中間合成音声情報を前記中間的な声質の合成音声に変換して出力する音声出力手段と
を備えることを特徴とする音声合成装置。
【請求項２】
前記モーフィング手段は、前記音声出力手段から出力される合成音声の声質がその出力中に連続的に変化するように、前記第１及び第２の合成音声情報の前記中間合成音声情報に対して寄与する割合を変化させる
ことを特徴とする請求項１記載の音声合成装置。
【請求項３】
前記音声情報生成手段は、前記第１及び第２の合成音声情報を音声生成モデルの特徴パラメタ値列として生成し、
前記モーフィング手段は、前記第１及び第２の合成音声情報の互いに対応するパラメタの中間値を計算することで、前記中間合成音声情報を生成する
ことを特徴とする請求項１又は２記載の音声合成装置。
【請求項４】
前記音声情報生成手段は、前記第１及び第２の合成音声情報を周波数スペクトルとして生成し、
前記モーフィング手段は、前記第１及び第２の合成音声情報の周波数スペクトルの中間的な周波数スペクトルを前記中間合成音声情報として生成する
ことを特徴とする請求項１又は２記載の音声合成装置。
【請求項５】
前記音声情報生成手段は、前記第１及び第２の合成音声情報を音声波形として生成し、
前記モーフィング手段は、前記第１及び第２の合成音声情報の音声波形の中間的な音声波形を前記中間合成音声情報として生成する
ことを特徴とする請求項１又は２記載の音声合成装置。
【請求項６】
前記記憶手段は、前記第１及び第２の音声素片情報のそれぞれにより示される各音声素片における基準を示す内容の特徴情報を、前記第１及び第２の音声素片情報のそれぞれに含めて記憶しており、
前記音声情報生成手段は、前記第１及び第２の合成音声情報を、それぞれに前記特徴情報を含めて生成し、
前記モーフィング手段は、前記第１及び第２の合成音声情報を、それぞれに含まれる前記特徴情報によって示される基準を用いて整合した上で前記中間合成音声情報を生成する
ことを特徴とする請求項１～５の何れか１項に記載の音声合成装置。
【請求項７】
前記基準は、前記第１及び第２の音声素片情報のそれぞれにより示される各音声素片の音響的特徴の変化点である

ことを特徴とする請求項６記載の音声合成装置。
【請求項８】
前記音響的特徴の変化点は、前記第１及び第２の音声素片情報のそれぞれに示される各音声素片をＨＭＭ（Hidden Markov Model）で表した最尤経路上の状態遷移点であって、
前記モーフィング手段は、前記第１及び第２の合成音声情報を、前記状態遷移点を用いて時間軸上で整合した上で前記中間合成音声情報を生成する
ことを特徴とする請求項７記載の音声合成装置。
【請求項９】
前記基準は、前記第１及び第２の音声素片情報のそれぞれにより示される各音声素片の、周波数スペクトルに現れる特徴的な周波数及び時間の関数であって、
前記モーフィング手段は、前記第１及び第２の合成音声情報を、前記関数を用いて周波数軸上で整合した上で前記中間合成音声情報を生成する
ことを特徴とする請求項６記載の音声合成装置。
【請求項１０】
前記音声合成装置は、さらに、
前記第１の声質に対応する画像を示す第１の画像情報、及び前記第２の声質に対応する画像を示す第２の画像情報を予め記憶している画像記憶手段と、
前記第１及び第２の画像情報のそれぞれにより示される画像の中間的な画像であって、前記中間合成音声情報の声質に対応する画像を示す中間画像情報を、前記第１及び第２の画像情報から生成する画像モーフィング手段と、
前記画像モーフィング手段により生成された中間画像情報を取得して、前記中間画像情報により示される画像を、前記音声出力手段から出力される合成音声に同期させて表示する表示手段と
を備えることを特徴とする請求項１～９の何れか１項に記載の音声合成装置。
【請求項１１】
前記第１の画像情報は前記第１の声質に対応する顔画像を示し、前記第２の画像情報は前記第２の声質に対応する顔画像を示す
ことを特徴とする請求項１０記載の音声合成装置。
【請求項１２】
前記音声合成装置は、さらに、
前記第１及び第２の声質を示す固定点、及びユーザの操作に基いて移動する移動点をそれぞれＮ次元（Ｎは自然数）の座標上に配置して表し、前記固定点及び移動点の配置に基いて、前記第１及び第２の合成音声情報の前記中間合成音声情報に対して寄与する割合を導出し、導出した割合を前記モーフィング手段に指示する指定手段を備え、
前記モーフィング手段は、前記指定手段により指定された割合に応じて、前記中間合成音声情報を生成する
ことを特徴とする請求項１～１１の何れか１項に記載の音声合成装置。
【請求項１３】
前記音声情報生成手段は、
前記第１及び第２の合成音声情報のそれぞれを順次生成する
ことを特徴とする請求項１～１２の何れか１項に記載の音声合成装置。
【請求項１４】
前記音声情報生成手段は、
前記第１及び第２の合成音声情報のそれぞれを並列に生成する
ことを特徴とする請求項１～１２の何れか１項に記載の音声合成装置。
【請求項１５】
第１の声質に属する複数の音声素片に関する第１の音声素片情報、及び前記第１の声質と異なる第２の声質に属する複数の音声素片に関する第２の音声素片情報を予め記憶しているメモリを用いることで、合成音声を生成して出力する音声合成方法であって、
テキストデータを取得するテキスト取得ステップと、

前記メモリの第１の音声素片情報から、前記テキストデータに含まれる文字に対応した前記第１の声質の合成音声を示す第１の合成音声情報を生成し、前記メモリの第２の音声素片情報から、前記テキストデータに含まれる文字に対応した前記第２の声質の合成音声を示す第２の合成音声情報を生成する音声情報生成ステップと、

前記音声情報生成ステップで生成された前記第１及び第２の合成音声情報から、前記テキストデータに含まれる文字に対応した、前記第１及び第２の声質の中間的な声質の合成音声を示す中間合成音声情報を生成するモーフィングステップと、

前記モーフィングステップで生成された前記中間合成音声情報を前記中間的な声質の合成音声に変換して出力する音声出力ステップと

を含むことを特徴とする音声合成方法。

【請求項１６】

前記モーフィングステップでは、前記音声出力ステップで出力される合成音声の声質がその出力中に連続的に変化するように、前記第１及び第２の合成音声情報の前記中間合成音声情報に対して寄与する割合を変化させる

ことを特徴とする請求項１５記載の音声合成方法。

【請求項１７】

前記音声情報生成ステップでは、前記第１及び第２の合成音声情報を音声生成モデルの特徴パラメタ値列として生成し、

前記モーフィングステップでは、前記第１及び第２の合成音声情報の互いに対応するパラメタの中間値を計算することで、前記中間合成音声情報を生成する

ことを特徴とする請求項１５又は１６記載の音声合成方法。

【請求項１８】

前記音声合成方法は、さらに、

前記第１の声質に対応する画像を示す第１の画像情報、及び前記第２の声質に対応する画像を示す第２の画像情報を予め記憶している画像メモリを用い、

前記第１及び第２の画像情報のそれぞれにより示される画像の中間的な画像であって、前記中間合成音声情報の声質に対応する画像を示す中間画像情報を、前記画像メモリの第１及び第２の画像情報から生成する画像モーフィングステップと、

前記画像モーフィングステップで生成された中間画像情報により示される画像を、前記音声出力ステップで出力される合成音声に同期させて表示する表示ステップと

を含むことを特徴とする請求項１５～１７の何れか１項に記載の音声合成方法。

【請求項１９】

前記第１の画像情報は前記第１の声質に対応する顔画像を示し、前記第２の画像情報は前記第２の声質に対応する顔画像を示す

ことを特徴とする請求項１８記載の音声合成方法。

【請求項２０】

請求項１５～１９の何れか１項に記載の音声合成方法に含まれるステップをコンピュータに実行させるプログラム。

【書類名】明細書
【発明の名称】音声合成装置
【技術分野】
【０００１】
本発明は、合成音声を生成して出力する音声合成装置に関する。
【背景技術】
【０００２】
従来より、所望の合成音声を生成して出力する音声合成装置が提供されている（例えば、特許文献１及び特許文献２参照。）。
特許文献１の音声合成装置は、それぞれ声質の異なる複数の音声素片データベースを備え、これらの音声素片データベースを切り替えて用いることにより、所望の合成音声を生成して出力する。
【０００３】
また、特許文献２の音声合成装置（音声変形装置）は、音声分析結果のスペクトルパラメタを変換することにより、所望の合成音声を生成して出力する。
【特許文献１】特開平７－３１９４９５号公報
【特許文献２】特開２０００－３３０５８２号公報
【発明の開示】
【発明が解決しようとする課題】
【０００４】
しかしながら、上記特許文献１及び特許文献２の音声合成装置では、声質変換の自由度が狭かったり、音質の調整が非常に困難であるという問題がある。
即ち、特許文献１では、合成音声の声質が予め設定された声質に限られ、その予め設定された声質間の連続的な変化を表現することができない。
また、特許文献２では、特に音声の基本周波数や音源パワーなどの特徴パラメタのダイナミックレンジを大きくしてしまうと音質に破綻が生じてしまい、良い音質を維持するのが困難となる。
【０００５】
そこで、本発明は、このような問題に鑑みてなされたものであって、声質の自由度が広く良い音質の合成音声をテキストデータから生成する音声合成装置を提供することを目的とする。
【課題を解決するための手段】
【０００６】
上記目的を達成するために、本発明に係る音声合成装置は、第１の声質に属する複数の音声素片に関する第１の音声素片情報、及び前記第１の声質と異なる第２の声質に属する複数の音声素片に関する第２の音声素片情報を予め記憶している記憶手段と、テキストデータを取得するとともに、前記記憶手段の第１の音声素片情報から、前記テキストデータに含まれる文字に対応した前記第１の声質の合成音声を示す第１の合成音声情報を生成し、前記記憶手段の第２の音声素片情報から、前記テキストデータに含まれる文字に対応した前記第２の声質の合成音声を示す第２の合成音声情報を生成する音声情報生成手段と、前記音声情報生成手段により生成された前記第１及び第２の合成音声情報から、前記テキストデータに含まれる文字に対応した、前記第１及び第２の声質の中間的な声質の合成音声を示す中間合成音声情報を生成するモーフィング手段と、前記モーフィング手段によって生成された前記中間合成音声情報を前記中間的な声質の合成音声に変換して出力する音声出力手段とを備えることを特徴とする。例えば、前記音声情報生成手段は、前記第１及び第２の合成音声情報を音声生成モデルの特徴パラメタ値列として生成し、前記モーフィング手段は、前記第１及び第２の合成音声情報の互いに対応するパラメタの中間値を計算することで、前記中間合成音声情報を生成する。
【０００７】
これにより、第１の声質に対する第１の音声素片情報、及び第２の声質に対する第２の

音声素片情報だけを記憶手段に予め記憶させておけば、第１及び第２の声質の中間的な声質の合成音声が出力されるため、記憶手段に予め記憶させておく内容の声質に限定されずに声質の自由度を広めることができる。また、第１及び第２の声質を有する第１及び第２の合成音声情報を基礎に中間合成音声情報が生成されるため、従来例のように基本周波数などのパラメタのダイナミックレンジを大きくしすぎるような処理がなされず、合成音声の音質を良い状態に維持することができる。また、本発明に係る音声合成装置は、テキストデータを取得して、そこに含まれる文字列に応じた合成音声を出力するため、ユーザに対する使い勝手を向上することができる。

【０００８】

ここで、前記モーフィング手段は、前記音声出力手段から出力される合成音声の声質がその出力中に連続的に変化するように、前記第１及び第２の合成音声情報の前記中間合成音声情報に対して寄与する割合を変化させることを特徴としても良い。

これにより、合成音声の出力中にその合成音声の声質が連続的に変化するため、例えば、平常声から怒り声に連続的に変化するような合成音声を出力することができる。

【０００９】

また、前記記憶手段は、前記第１及び第２の音声素片情報のそれぞれにより示される各音声素片における基準を示す内容の特徴情報を、前記第１及び第２の音声素片情報のそれぞれに含めて記憶しており、前記音声情報生成手段は、前記第１及び第２の合成音声情報を、それぞれに前記特徴情報を含めて生成し、前記モーフィング手段は、前記第１及び第２の合成音声情報を、それぞれに含まれる前記特徴情報によって示される基準を用いて整合した上で前記中間合成音声情報を生成することを特徴としても良い。例えば、前記基準は、前記第１及び第２の音声素片情報のそれぞれにより示される各音声素片の音響的特徴の変化点である。

【００１０】

これにより、モーフィング手段による中間合成音声情報の生成に、第１及び第２の合成音声情報が上述の基準を用いて整合されるため、例えば第１及び第２の合成音声情報をパターンマッチングなどによって整合するような場合と比べ、迅速に整合を図って中間合成音声情報を生成することができ、その結果、処理速度を向上することができる。

【００１１】

また、前記音声合成装置は、さらに、前記第１の声質に対応する画像を示す第１の画像情報、及び前記第２の声質に対応する画像を示す第２の画像情報を予め記憶している画像記憶手段と、前記第１及び第２の画像情報のそれぞれにより示される画像の中間的な画像であって、前記中間合成音声情報の声質に対応する画像を示す中間画像情報を、前記第１及び第２の画像情報から生成する画像モーフィング手段と、前記画像モーフィング手段により生成された中間画像情報を取得して、前記中間画像情報により示される画像を、前記音声出力手段から出力される合成音声に同期させて表示する表示手段とを備えることを特徴としても良い。例えば、前記第１の画像情報は前記第１の声質に対応する顔画像を示し、前記第２の画像情報は前記第２の声質に対応する顔画像を示す。

【００１２】

これにより、第１及び第２の声質の中間的な声質に対応する顔画像が、その中間的な声質の合成音声の出力と同期して表示されるため、合成音声の声質を顔画像の表情からもユーザに伝えることができ、表現力の向上を図ることができる。

【００１３】

ここで、前記音声情報生成手段は、前記第１及び第２の合成音声情報のそれぞれを順次生成することを特徴としても良い。

これにより、音声情報生成手段の単位時間あたりの処理負担を軽減することができ、音声情報生成手段の構成を簡単にすることができる。その結果、装置全体を小型化することができるとともに、コスト低減を図ることができる。

【００１４】

また、前記音声情報生成手段は、前記第１及び第２の合成音声情報のそれぞれを並列に

生成することを特徴としても良い。
これにより、第１及び第２の合成音声情報を迅速に生成することができ、その結果、テキストデータの取得から合成音声の出力までの時間を短縮することができる。
なお、本発明は、上述の音声合成装置の合成音声を生成して出力する方法やプログラム、そのプログラムを格納する記憶媒体としても実現することができる。
【発明の効果】
【００１５】
本発明の音声合成装置では、声質の自由度が広く良い音質の合成音声をテキストデータから生成することができるという効果を奏する。
【発明を実施するための最良の形態】
【００１６】
以下、本発明の実施の形態について図面を用いて詳細に説明する。
（実施の形態１）
図１は、本発明の実施の形態１に係る音声合成装置の構成を示す構成図である。
本実施の形態の音声合成装置は、声質の自由度が広く良い音質の合成音声をテキストデータから生成するものであって、複数の音声素片（音素）に関する音声素片データを蓄積する複数の音声合成ＤＢ１０１ａ～１０１ｚと、１つの音声合成ＤＢに蓄積された音声素片データを用いることにより、テキスト１０に示される文字列に対応する音声合成パラメタ値列１１を生成する複数の音声合成部（音声情報生成手段）１０３と、ユーザによる操作に基づいて声質を指定する声質指定部１０４と、複数の音声合成部１０３により生成された音声合成パラメタ値列１１を用いて音声モーフィング処理を行い、中間的合成音波形データ１２を出力する音声モーフィング部１０５と、中間的合成音波形データ１２に基づいて合成音声を出力するスピーカ１０７とを備えている。
【００１７】
音声合成ＤＢ１０１ａ～１０１ｚのそれぞれが蓄積する音声素片データの示す声質は異なっている。例えば、音声合成ＤＢ１０１ａには、笑っている声質の音声素片データが蓄積され、音声合成ＤＢ１０１ｚには、怒っている声質の音声素片データが蓄積されている。また、本実施の形態における音声素片データは、音声生成モデルの特徴パラメタ値列の形式で表現されている。さらに、蓄積される各音声素片データには、これらのデータにより示される各音声素片の開始及び終了の時刻と、音響的特徴の変化点の時刻とを示すラベル情報が付されている。
【００１８】
複数の音声合成部１０３は、それぞれ上述の音声合成ＤＢと一対一に対応付けられている。このような音声合成部１０３の動作について図２を参照して説明する。
図２は、音声合成部１０３の動作を説明するための説明図である。
音声合成部１０３は、図２に示すように、言語処理部１０３ａと素片結合部１０３ｂとを備えている。
【００１９】
言語処理部１０３ａは、テキスト１０を取得して、テキスト１０に示される文字列を音素情報１０ａに変換する。音素情報１０ａは、テキスト１０に示される文字列が音素列の形で表現されたもので、他にアクセント位置情報や音素継続長情報など、素片選択・結合・変形に必要な情報を含んでもよい。
素片結合部１０３ｂは、対応付けられた音声合成ＤＢの音声素片データから適切な音声素片に関する部分を抜き出して、抜き出した部分の結合と変形を行うことにより、言語処理部１０３ａにより出力される音素情報１０ａに対応する音声合成パラメタ値列１１を生成する。音声合成パラメタ値列１１は、実際の音声波形を生成するために必要となる十分な情報を含んだ複数の特徴パラメタの値が配列されたものである。例えば、音声合成パラメタ値列１１は、時系列に沿った各音声分析合成フレームごとに、図２に示すような、５つの特徴パラメタを含んで構成される。５つの特徴パラメタとは、音声の基本周波数Ｆ０と、第一フォルマントＦ１と、第二フォルマントＦ２と、音声分析合成フレーム継続長Ｆ

Ｒと、音源強度ＰＷとである。また、上述のように音声素片データにはラベル情報が付されているので、このように生成される音声合成パラメタ値列１１にもラベル情報が付されている。

【００２０】

声質指定部１０４は、ユーザによる操作に基づき、何れの音声合成パラメタ値列１１を用い、その音声合成パラメタ値列１１に対してどのような割合で音声モーフィング処理を行うかを音声モーフィング部１０５に指示する。さらに、声質指定部１０４はその割合を時系列に沿って変化させる。このような声質指定部１０４は、例えばパーソナルコンピュータなどから構成され、ユーザにより操作された結果を表示するディスプレイを備えている。

【００２１】

図３は、声質指定部１０４のディスプレイが表示する画面の一例を示す画面表示図である。

ディスプレイには、音声合成部ＤＢ１０１ａ～１０１ｚの声質を示す複数の声質アイコンが表示されている。なお図３では、複数の声質アイコンのうち、声質Ａの声質アイコン１０４Ａと、声質Ｂの声質アイコン１０４Ｂと、声質Ｚの声質アイコン１０４Ｚとを示す。このような複数の声質アイコンは、それぞれの示す声質が似ているものほど互いに近寄るように配置され、似ていないものほど互いに離れるように配置される。

【００２２】

ここで、声質指定部１０４は、このようなディスプレイ上に、ユーザによる操作に応じて移動可能な指定アイコン１０４ｉを表示する。

声質指定部１０４は、ユーザによって配置された指定アイコン１０４ｉから近い声質アイコンを調べ、例えば声質アイコン１０４Ａ，１０４Ｂ，１０４Ｚを特定すると、声質Ａの音声合成パラメタ値列１１と、声質Ｂの音声合成パラメタ値列１１と、声質Ｚの音声合成パラメタ値列１１とを用いることを、音声モーフィング部１０５に指示する。さらに、声質指定部１０４は、各声質アイコン１０４Ａ，１０４Ｂ，１０４Ｚ及び指定アイコン１０４ｉの相対的な配置に対応する割合を、音声モーフィング部１０５に指示する。

【００２３】

即ち、声質指定部１０４は、指定アイコン１０４ｉから各声質アイコン１０４Ａ，１０４Ｂ，１０４Ｚまでの距離を調べ、それらの距離に応じた割合を指示する。

又は、声質指定部１０４は、まず、声質Ａと声質Ｚの中間的な声質（テンポラリ声質）を生成するための割合を求め、次に、そのテンポラリ声質と声質Ｂとから、指定アイコン１０４ｉで示される声質を生成するための割合を求め、これらの割合を指示する。具体的に、声質指定部１０４は、声質アイコン１０４Ａ及び声質アイコン１０４Ｚを結ぶ直線と、声質アイコン１０４Ｂ及び指定アイコン１０４ｉを結ぶ直線とを算出し、これらの直線の交点の位置１０４ｔを特定する。この位置１０４ｔにより示される声質が上述のテンポラリ声質である。そして、声質指定部１０４は、位置１０４ｔから各声質アイコン１０４Ａ，１０４Ｚまでの距離の割合を求める。次に、声質指定部１０４は、指定アイコン１０４ｉから声質アイコン１０４Ｂ及び位置１０４ｔまでの距離の割合を求め、このように求めた２つの割合を指示する。

【００２４】

このような声質指定部１０４を操作することにより、ユーザは、スピーカ１０７から出力させようとする合成音声の声質の、予め設定された声質に対する類似度を容易に入力することができる。そこでユーザは、例えば声質Ａに近い合成音声をスピーカ１０７から出力させたいときには、指定アイコン１０４ｉが声質アイコン１０４Ａに近づくように声質指定部１０４を操作する。

また、声質指定部１０４は、ユーザからの操作に応じて、上述のような割合を時系列に沿って連続的に変化させる。

【００２５】

図４は、声質指定部１０４のディスプレイが表示する他の画面の一例を示す画面表示図

である。
声質指定部１０４は、図４に示すように、ユーザによる操作に応じて、ディスプレイ上に３つのアイコン２１，２２，２３を配置し、アイコン２１からアイコン２２を通ってアイコン２３に到達するような軌跡を特定する。そして、声質指定部１０４は、その軌跡に沿って指定アイコン１０４ｉが移動するように、上述の割合を時系列に沿って連続的に変化させる。例えば、声質指定部１０４は、その軌跡の長さをＬとすると、毎秒０．０１×Ｌの速度で指定アイコン１０４ｉが移動するように、その割合を変化させる。
音声モーフィング部１０５は、上述のような声質指定部１０４により指定された音声合成パラメタ値列１１と割合とから、音声モーフィング処理を行う。
【００２６】
図５は、音声モーフィング部１０５の処理動作を説明するための説明図である。
音声モーフィング部１０５は、図５に示すように、パラメタ中間値計算部１０５ａと、波形生成部１０５ｂとを備えている。
【００２７】
パラメタ中間値計算部１０５ａは、声質指定部１０４により指定された少なくとも２つの音声合成パラメタ値列１１と割合とを特定し、それらの音声合成パラメタ値列１１から、互いに対応する音声分析合成フレーム間ごとに、その割合に応じた中間的音声合成パラメタ値列１３を生成する。
例えば、パラメタ中間値計算部１０５ａは、声質指定部１０４の指定に基づいて、声質Ａの音声合成パラメタ値列１１と、声質Ｚの音声合成パラメタ値列１１と、割合５０：５０とを特定すると、まず、その声質Ａの音声合成パラメタ値列１１と、声質Ｚの音声合成パラメタ値列１１とを、それぞれに対応する音声合成部１０３から取得する。そして、パラメタ中間値計算部１０５ａは、互いに対応する音声分析合成フレームにおいて、声質Ａの音声合成パラメタ値列１１に含まれる各特徴パラメタと、声質Ｚの音声合成パラメタ値列１１に含まれる各特徴パラメタとの中間値を５０：５０の割合で算出し、その算出結果を中間的音声合成パラメタ値列１３として生成する。具体的に、互いに対応する音声分析合成フレームにおいて、声質Ａの音声合成パラメタ値列１１の基本周波数Ｆ０の値が３００であり、声質Ｚの音声合成パラメタ値列１１の基本周波数Ｆ０の値が２８０である場合には、パラメタ中間値計算部１０５ａは、当該音声分析合成フレームでの基本周波数Ｆ０が２９０となる中間的音声合成パラメタ値列１３を生成する。
【００２８】
また、図３を用いて説明したように、声質指定部１０４により、声質Ａの音声合成パラメタ値列１１と、声質Ｂの音声合成パラメタ値列１１と、声質Ｚの音声合成パラメタ値列１１とが指定され、さらに、声質Ａと声質Ｚの中間的なテンポラリ声質を生成するための割合（例えば３：７）と、そのテンポラリ声質と声質Ｂとから指定アイコン１０４ｉで示される声質を生成するための割合（例えば９：１）とが指定され場合には、音声モーフィング部１０５は、まず、声質Ａの音声合成パラメタ値列１１と、声質Ｚの音声合成パラメタ値列１１とを用いて、３：７の割合に応じた音声モーフィング処理を行う。これにより、テンポラリ声質に対応する音声合成パラメタ値列が生成される。さらに、音声モーフィング部１０５は、先に生成した音声合成パラメタ値列と、声質Ｂの音声合成パラメタ値列１１とを用いて、９：１の割合に応じた音声モーフィング処理を行う。これにより、指定アイコン１０４ｉに対応する中間的音声合成パラメタ値列１３が生成される。ここで、上述の３：７の割合に応じた音声モーフィング処理とは、声質Ａの音声合成パラメタ値列１１を３／（３＋７）だけ声質Ｚの音声合成パラメタ値列１１に近づける処理であり、逆に、声質Ｚの音声合成パラメタ値列１１を７／（３＋７）だけ声質Ａの音声合成パラメタ値列１１に近づける処理をいう。この結果、生成される音声合成パラメタ値列は、声質Ｚの音声合成パラメタ値列１１よりも、声質Ａの音声合成パラメタ値列１１に類似することとなる。
【００２９】
波形生成部１０５ｂは、パラメタ中間値計算部１０５ａにより生成された中間的音声合

成パラメタ値列１３を取得して、その中間的音声合成パラメタ値列１３に応じた中間的合成音波形データ１２を生成し、スピーカ１０７に対して出力する。

これにより、スピーカ１０７からは、中間的音声合成パラメタ値列１３に応じた合成音声が出力される。即ち、予め設定された複数の声質の中間的な声質の合成音声がスピーカ１０７から出力される。

【００３０】

ここで、一般に複数の音声合成パラメタ値列１１に含まれる音声分析合成フレームの総数はそれぞれ異なるため、パラメタ中間値計算部１０５ａは、上述のように互いに異なる声質の音声合成パラメタ値列１１を用いて音声モーフィング処理を行うときには、音声分析合成フレーム間の対応付けを行うために時間軸アライメントを行う。

即ちパラメタ中間値計算部１０５ａは、音声合成パラメタ値列１１に付されたラベル情報に基づいて、これらの音声合成パラメタ値列１１の時間軸上の整合を図る。

【００３１】

ラベル情報は、前述のように各音声素片の開始及び終了の時刻と、音響的特徴の変化点の時刻とを示す。音響的特徴の変化点は、例えば、音声素片に対応する不特定話者ＨＭＭ音素モデルにより示される最尤パスの状態遷移点である。

図６は、音声素片とＨＭＭ音素モデルの一例を示す例示図である。

例えば、図６に示すように、所定の音声素片３０を不特定話者ＨＭＭ音素モデル（以下、音素モデルと略す）３１で認識した場合、その音素モデル３１は、開始状態（S_0）と終了状態（S_E）を含めて４つの状態（S_0，S_1，S_2，S_E）で構成される。ここで、最尤パスの形状３２は、時刻４から５において、状態Ｓ１から状態Ｓ２への状態遷移を有する。つまり、音声合成ＤＢ１０１ａ～１０１ｚに格納されている音声素片データの音声素片３０に対応する部分には、この音声素片３０の開始時刻１、終了時刻Ｎ、及び音響的特徴の変化点の時刻５を示すラベル情報が付されている。

【００３２】

したがって、パラメタ中間値計算部１０５ａは、そのラベル情報に示される開始時刻１、終了時刻Ｎ、及び音響的特徴の変換点の時刻５に基づいて、時間軸の伸縮処理を行う。即ち、パラメタ中間値計算部１０５ａは、取得した各音声合成パラメタ値列１１に対して、ラベル情報により示される時刻が一致するように、その時刻間を線形に伸縮する。

これにより、パラメタ中間値計算部１０５ａは、各音声合成パラメタ値列１１に対して、それぞれの音声分析合成フレームの対応付けを行うことができる。つまり、時間軸アライメントを行うことができる。また、このように本実施の形態ではラベル情報を用いて時間軸アライメントを行うことにより、例えば各音声合成パラメタ値列１１のパターンマッチングなどにより時間軸アライメントを行う場合と比べて、迅速に時間軸アライメントを実行することができる。

【００３３】

以上のように本実施の形態では、パラメタ中間値計算部１０５ａが、声質指定部１０４から指示された複数の音声合成パラメタ値列１１に対して、声質指定部１０４から指定された割合に応じた音声モーフィング処理を実行するため、合成音声の声質の自由度を広めることができる。

例えば、図３に示す声質指定部１０４のディスプレイ上で、ユーザが声質指定部１０４を操作することにより指定アイコン１０４ｉを声質アイコン１０４Ａ、声質アイコン１０４Ｂ及び声質アイコン１０４Ｚに近づければ、音声モーフィング部１０５は、声質Ａの音声合成ＤＢ１０１ａに基づいて音声合成部１０３により生成された音声合成パラメタ値列１１と、声質Ｂの音声合成ＤＢ１０１ｂに基づいて音声合成部１０３により生成された音声合成パラメタ値列１１と、声質Ｚの音声合成ＤＢ１０１ｚに基づいて音声合成部１０３により生成された音声合成パラメタ値列１１とを用いて、それぞれを同じ割合で音声モーフィング処理する。その結果、スピーカ１０７から出力される合成音声を、声質Ａと声質Ｂと声質Ｃとの中間的な声質にすることができる。また、ユーザが声質指定部１０４を操作することにより指定アイコン１０４ｉを声質アイコン１０４Ａに近づければ、スピーカ

１０７から出力される合成音声の声質を声質Ａに近づけることができる。

【００３４】

また、本実施の形態の声質指定部１０４は、ユーザによる操作に応じてその割合を時系列に沿って変化させるため、スピーカ１０７から出力される合成音声の声質を時系列に沿ってなめらかに変化させることができる。例えば、図４で説明したように、声質指定部１０４が、毎秒０．０１×Ｌの速度で軌跡上を指定アイコン１０４ｉが移動するように割合を変化させた場合には、１００秒間声質がなめらかに変化し続けるような合成音声がスピーカ１０７から出力される。

【００３５】

これによって、例えば「喋り始めは冷静だが、喋りながら段々怒っていく」というような、従来は不可能だった、表現力の高い音声合成装置が実現できる。また、合成音声の声質を１発声の中で連続的に変化させることもできる。

さらに、本実施の形態では、音声モーフィング処理を行うため、従来例のように声質に破綻が起こることがなく合成音声の品質を維持することができる。

【００３６】

なお、本実施の形態では、複数の音声合成部１０３のそれぞれに音素情報１０ａ及び音声合成パラメタ値列１１を生成させたが、音声モーフィング処理に必要となる声質に対応する音素情報１０ａが何れも同じであるときには、１つの音声合成部１０３の言語処理部１０３ａにのみ音素情報１０ａを生成させ、その音素情報１０ａから音声合成パラメタ値列１１を生成する処理を、複数の音声合成部１０３の素片結合部１０３ｂにさせても良い。

【００３７】

（変形例）

ここで、本実施の形態における音声合成部に関する変形例について説明する。

図７は、本変形例に係る音声合成装置の構成を示す構成図である。

本変形例に係る音声合成装置は、互いに異なる声質の音声合成パラメタ値列１１を生成する１つの音声合成部１０３ｃを備える。

【００３８】

この音声合成部１０３ｃは、テキスト１０を取得して、テキスト１０に示される文字列を音素情報１０ａに変換した後、複数の音声合成ＤＢ１０１ａ～１０１ｚを順番に切り替えて参照ことで、その音素情報１０ａに対応する複数の声質の音声合成パラメタ値列１１を順次生成する。

音声モーフィング部１０５は、必要な音声合成パラメタ値列１１が生成されるまで待機し、その後、上述と同様の方法で中間的合成音波形データ１２を生成する。

【００３９】

なお、上述のような場合、声質指定部１０４は、音声合成部１０３ｃに指示して、音声モーフィング部１０５が必要とする音声合成パラメタ値列１１のみを生成させることで、音声モーフィング部１０５の待機時間を短くすることができる。

このように本変形例では、音声合成部１０３ｃを１つだけ備えることにより、音声合成装置全体の小型化並びにコスト低減を図ることができる。

【００４０】

（実施の形態２）

図８は、本発明の実施の形態２に係る音声合成装置の構成を示す構成図である。

本実施の形態の音声合成装置は、実施の形態１の音声合成パラメタ値列１１の代わりに周波数スペクトルを用い、この周波数スペクトルによる音声モーフィング処理を行う。

このような音声合成装置は、複数の音声素片に関する音声素片データを蓄積する複数の音声合成ＤＢ２０１ａ～２０１ｚと、１つの音声合成ＤＢに蓄積された音声素片データを用いることにより、テキスト１０に示される文字列に対応する合成音スペクトル４１を生成する複数の音声合成部２０３と、ユーザによる操作に基づいて声質を指定する声質指定部１０４と、複数の音声合成部２０３により生成された合成音スペクトル４１を用いて音

声モーフィング処理を行い、中間的合成音波形データ１２を出力する音声モーフィング部２０５と、中間的合成音波形データ１２に基づいて合成音声を出力するスピーカ１０７とを備えている。

【００４１】

複数の音声合成ＤＢ２０１ａ～２０１ｚのそれぞれが蓄積する音声素片データの示す声質は、実施の形態１の音声合成ＤＢ１０１ａ～１０１ｚと同様、異っている。また、本実施の形態における音声素片データは、周波数スペクトルの形式で表現されている。

複数の音声合成部２０３は、それぞれ上述の音声合成ＤＢと一対一に対応付けられている。そして、各音声合成部２０３は、テキスト１０を取得して、テキスト１０に示される文字列を音素情報に変換する。さらに、音声合成部２０３は、対応付けられた音声合成ＤＢの音声素片データから適切な音声素片に関する部分を抜き出して、抜き出した部分の結合と変形を行うことにより、先に生成した音素情報に対応する周波数スペクトルたる合成音スペクトル４１を生成する。このような合成音スペクトル４１は、音声のフーリエ解析結果の形式であっても良く、音声のケプストラムパラメタ値を時系列的に並べた形式であっても良い。

【００４２】

声質指定部１０４は、実施の形態１と同様、ユーザによる操作に基づき、何れの合成音スペクトル４１を用い、その合成音スペクトル４１に対してどのような割合で音声モーフィング処理を行うかを音声モーフィング部２０５に指示する。さらに、声質指定部１０４はその割合を時系列に沿って変化させる。

本実施の形態における音声モーフィング部２０５は、複数の音声合成部２０３から出力される合成音スペクトル４１を取得して、その中間的性質を持つ合成音スペクトルを生成し、さらに、その中間的性質の合成音スペクトルを中間的合成音波形データ１２に変形して出力する。

【００４３】

図９は、本実施の形態における音声モーフィング部２０５の処理動作を説明するための説明図である。

音声モーフィング部２０５は、図９に示すように、スペクトルモーフィング部２０５ａと、波形生成部２０５ｂとを備えている。

スペクトルモーフィング部２０５ａは、声質指定部１０４により指定された少なくとも２つの合成音スペクトル４１と割合とを特定し、それらの合成音スペクトル４１から、その割合に応じた中間的合成音スペクトル４２を生成する。

【００４４】

即ち、スペクトルモーフィング部２０５ａは、複数の合成音スペクトル４１から、声質指定部１０４により指定された２つ以上の合成音スペクトル４１を選択する。そして、スペクトルモーフィング部２０５ａは、それら合成音スペクトル４１の形状の特徴を示すフォルマント形状５０を抽出して、そのフォルマント形状５０ができるだけ一致するような変形を各合成音スペクトル４１に加えた後、各合成音スペクトル４１の重ね合わせを行う。なお、上述の合成音スペクトル４１の形状の特徴は、フォルマント形状でなくても良く、例えばある程度以上強く現れていて、かつその軌跡が連続的に追えるものであれば良い。図９に示されるように、フォルマント形状５０は、声質Ａの合成音スペクトル４１及び声質Ｚの合成音スペクトル４１のそれぞれについてスペクトル形状の特徴を模式的に表すものである。

【００４５】

具体的に、スペクトルモーフィング部２０５ａは、声質指定部１０４からの指定に基づき、声質Ａ及び声質Ｚの合成音スペクトル４１と４：６の割合とを特定すると、まず、その声質Ａの合成音スペクトル４１と声質Ｚの合成音スペクトル４１とを取得して、それらの合成音スペクトル４１からフォルマント形状５０を抽出する。次に、スペクトルモーフィング部２０５ａは、声質Ａの合成音スペクトル４１のフォルマント形状５０が声質Ｚの合成音スペクトル４１のフォルマント形状５０に４０％だけ近づくように、声質Ａの合成

音スペクトル４１を周波数軸及び時間軸上で伸縮処理する。さらに、スペクトルモーフィング部２０５ａは、声質Ｚの合成音スペクトル４１のフォルマント形状５０が声質Ａの合成音スペクトル４１のフォルマント形状５０に６０％だけ近づくように、声質Ｚの合成音スペクトル４１を周波数軸及び時間軸上で伸縮処理する。最後に、スペクトルモーフィング部２０５ａは、伸縮処理された声質Ａの合成音スペクトル４１のパワーを６０％にするとともに、伸縮処理された声質Ｚの合成音スペクトル４１のパワーを４０％にした上で、両合成音スペクトル４１を重ね合わせる。その結果、声質Ａの合成音スペクトル４１と声質Ｚの合成音スペクトル４１との音声モーフィング処理が４：６の割合で行われ、中間的合成音スペクトル４２が生成される。

【００４６】

このような、中間的合成音スペクトル４２を生成する音声モーフィング処理について、図１０～図１２を用いてより詳細に説明する。

図１０は、声質Ａ及び声質Ｚの合成音スペクトル４１と、それらに対応する短時間フーリエスペクトルとを示す図である。

スペクトルモーフィング部２０５ａは、声質Ａの合成音スペクトル４１と声質Ｚの合成音スペクトル４１との音声モーフィング処理を４：６の割合で行うときには、まず、上述のようにこれらの合成音スペクトル４１のフォルマント形状５０を互いに近づけるため、各合成音スペクトル４１同士の時間軸アライメントを行う。このような時間軸アライメントは、各合成音スペクトル４１のフォルマント形状５０同士のパターンマッチングを行うことにより実現される。なお、各合成音スペクトル４１もしくはフォルマント形状５０に関する他の特徴量を用いてパターンマッチングを行ってもよい。

【００４７】

即ち、スペクトルモーフィング部２０５ａは、図１０に示すように、両合成音スペクトル４１のそれぞれのフォルマント形状５０において、パターンが一致するフーリエスペクトル分析窓５１の部位で時刻が一致するように、両合成音スペクトル４１に対して時間軸上の伸縮を行う。これにより時間軸アライメントが実現される。

また、図１０に示すように、互いにパターンが一致するフーリエスペクトル分析窓５１のそれぞれの短時間フーリエスペクトル４１ａには、フォルマント形状５０の周波数５０ａ，５０ｂが互いに異なるように表示される。

【００４８】

そこで、時間軸アライメントの完了後、スペクトルモーフィング部２０５ａは、アライメントされた音声の各時刻において、フォルマント形状５０を基に、周波数軸上の伸縮処理を行う。即ち、スペクトルモーフィング部２０５ａは、各時刻における声質Ａ及び声質Ｂの短時間フーリエスペクトル４１ａにおいて周波数５０ａ，５０ｂが一致するように、両短時間フーリエスペクトル４１ａを周波数軸上で伸縮する。

【００４９】

図１１は、スペクトルモーフィング部２０５ａが両短時間フーリエスペクトル４１ａを周波数軸上で伸縮する様子を説明するための説明図である。

スペクトルモーフィング部２０５ａは、声質Ａの短時間フーリエスペクトル４１ａ上の周波数５０ａ，５０ｂが４０％だけ、声質Ｚの短時間フーリエスペクトル４１ａ上の周波数５０ａ，５０ｂに近付くように、声質Ａの短時間フーリエスペクトル４１ａを周波数軸上で伸縮し、中間的な短時間フーリエスペクトル４１ｂを生成する。これと同様に、スペクトルモーフィング部２０５ａは、声質Ｚの短時間フーリエスペクトル４１ａ上の周波数５０ａ，５０ｂが６０％だけ、声質Ａの短時間フーリエスペクトル４１ａ上の周波数５０ａ，５０ｂに近付くように、声質Ｚの短時間フーリエスペクトル４１ａを周波数軸上で伸縮し、中間的な短時間フーリエスペクトル４１ｂを生成する。その結果、中間的な両短時間フーリエスペクトル４１ｂにおいて、フォルマント形状５０の周波数は周波数ｆ１，ｆ２に揃えられた状態となる。

【００５０】

例えば、声質Ａの短時間フーリエスペクトル４１ａ上でフォルマント形状５０の周波数

５０ａ，５０ｂが５００Ｈｚ及び３０００Ｈｚであり、声質Ｚの短時間フーリエスペクトル４１ａ上でフォルマント形状５０の周波数５０ａ，５０ｂが４００Ｈｚ及び４０００Ｈｚであり、かつ各合成音のナイキスト周波数が１１０２５Ｈｚである場合を想定して説明する。スペクトルモーフィング部２０５ａは、まず、声質Ａの短時間フーリエスペクトル４１ａの帯域ｆ＝０～５００Ｈｚが０～（５００＋（４００－５００）＊０．４）Ｈｚとなるように、帯域ｆ＝５００～３０００Ｈｚが（５００＋（４００－５００）＊０．４）～（３０００＋（４０００－３０００）＊０．４）Ｈｚとなるように、帯域ｆ＝３０００～１１０２５Ｈｚが（３０００＋（４０００－３０００）＊０．４）～１１０２５Ｈｚとなるように、声質Ａの短時間フーリエスペクトル４１ａに対して周波数軸上の伸縮・移動を行う。これと同様に、スペクトルモーフィング部２０５ａは、声質Ｚの短時間フーリエスペクトル４１ａの帯域ｆ＝０～４００Ｈｚが０～（４００＋（５００－４００）＊０．６）Ｈｚとなるように、帯域ｆ＝４００～４０００Ｈｚが（４００＋（５００－４００）＊０．６）～（４０００＋（３０００－４０００）＊０．６）Ｈｚとなるように、帯域ｆ＝４０００～１１０２５Ｈｚが（４０００＋（３０００－４０００）＊０．６）～１１０２５Ｈｚとなるように、声質Ｚの短時間フーリエスペクトル４１ａに対して周波数軸上の伸縮・移動を行う。その伸縮・移動の結果により生成された２つの短時間フーリエスペクトル４１ｂにおいて、フォルマント形状５０の周波数は周波数ｆ１，ｆ２に揃えられた状態となる。

【００５１】

次に、スペクトルモーフィング部２０５ａは、このような周波数軸上の変形が行われた両短時間フーリエスペクトル４１ｂのパワーを変形する。即ち、スペクトルモーフィング部２０５ａは、声質Ａの短時間フーリエスペクトル４１ｂのパワーを６０％に変換し、声質Ｚの短時間フーリエスペクトル４１ｂのパワーを４０％に変換する。そして、スペクトルモーフィング部２０５ａは、上述のように、パワーが変換されたこれらの短時間フーリエスペクトルを重ね合わせる。

【００５２】

図１２は、パワーが変換された２つの短時間フーリエスペクトルを重ね合わせる様子を説明するための説明図である。

この図１２に示すように、スペクトルモーフィング部２０５ａは、パワーが変換された声質Ａの短時間フーリエスペクトル４１ｃと、同じくパワーが変換された声質Ｂの短時間フーリエスペクトル４１ｃとを重ね合わせ、新たな短時間フーリエスペクトル４１ｄを生成する。このとき、スペクトルモーフィング部２０５ａは、互いの短時間フーリエスペクトル４１ｃの上記周波数ｆ１，ｆ２を一致させた状態で、両短時間フーリエスペクトル４１ｃを重ね合わせる。

【００５３】

そして、スペクトルモーフィング部２０５ａは、上述のような短時間フーリエスペクトル４１ｄの生成を、両合成音スペクトル４１の時間軸アライメントされた時刻ごとに行う。その結果、声質Ａの合成音スペクトル４１と声質Ｚの合成音スペクトル４１との音声モーフィング処理が４：６の割合で行われ、中間的合成音スペクトル４２が生成されるのである。

【００５４】

音声モーフィング部２０５の波形生成部２０５ｂは、上述のようにスペクトルモーフィング部２０５ａにより生成された中間的合成音スペクトル４２を、中間的合成音波形データ１２に変換して、これをスピーカ１０７に出力する。その結果、スピーカ１０７から、中間的合成音スペクトル４２に対応する合成音声が出力される。

このように、本実施の形態においても、実施の形態１と同様、声質の自由度が広く良い音質の合成音声をテキスト１０から生成することができる。

【００５５】

（変形例）

ここで、本実施の形態におけるスペクトルモーフィング部の動作に関する変形例につい

て説明する。
本変形例に係るスペクトルモーフィング部は、上述のように合成音スペクトル４１からその形状の特徴を示すフォルマント形状５０を抽出して用いることなく、音声合成ＤＢに予め格納されたスプライン曲線の制御点の位置を読み出して、そのスプライン曲線をフォルマント形状５０の代わりに用いる。
【００５６】
即ち、各音声素片に対応するフォルマント形状５０を、周波数対時間の２次元平面上の複数のスプライン曲線と見なし、そのスプライン曲線の制御点の位置を予め音声合成ＤＢに格納しておく。
このように、本変形例に係るスペクトルモーフィング部は、合成音スペクトル４１からわざわざフォルマント形状５０を抽出することをせず、音声合成ＤＢに予め格納されている制御点の位置が示すスプライン曲線を用いて時間軸及び周波数軸上の変換処理を行うため、上記変換処理を迅速に行うことができる。
【００５７】
なお、上述のようなスプライン曲線の制御点の位置ではなくフォルマント形状５０そのものを、予め音声合成ＤＢ２０１ａ～２０１ｚに格納しておいても良い。
【００５８】
（実施の形態３）
図１３は、本発明の実施の形態３に係る音声合成装置の構成を示す構成図である。
本実施の形態の音声合成装置は、実施の形態１の音声合成パラメタ値列１１や、実施の形態２の合成音スペクトル４１の代わりに音声波形を用い、この音声波形による音声モーフィング処理を行う。
【００５９】
このような音声合成装置は、複数の音声素片に関する音声素片データを蓄積する複数の音声合成ＤＢ３０１ａ～３０１ｚと、１つの音声合成ＤＢに蓄積された音声素片データを用いることにより、テキスト１０に示される文字列に対応する合成音波形データ６１を生成する複数の音声合成部３０３と、ユーザによる操作に基づいて声質を指定する声質指定部１０４と、複数の音声合成部３０３により生成された合成音波形データ６１を用いて音声モーフィング処理を行い、中間的合成音波形データ１２を出力する音声モーフィング部３０５と、中間的合成音波形データ１２に基づいて合成音声を出力するスピーカ１０７とを備えている。
【００６０】
複数の音声合成ＤＢ３０１ａ～３０１ｚのそれぞれが蓄積する音声素片データの示す声質は、実施の形態１の音声合成ＤＢ１０１ａ～１０１ｚと同様、異なっている。また、本実施の形態における音声素片データは、音声波形の形式で表現されている。
複数の音声合成部３０３は、それぞれ上述の音声合成ＤＢと一対一に対応付けられている。そして、各音声合成部３０３は、テキスト１０を取得して、テキスト１０に示される文字列を音素情報に変換する。さらに、音声合成部３０３は、対応付けられた音声合成ＤＢの音声素片データから適切な音声素片に関する部分を抜き出して、抜き出した部分の結合と変形を行うことにより、先に生成した音素情報に対応する音声波形たる合成音波形データ６１を生成する。
【００６１】
声質指定部１０４は、実施の形態１と同様、ユーザによる操作に基づき、何れの合成音波形データ６１を用い、その合成音波形データ６１に対してどのような割合で音声モーフィング処理を行うかを音声モーフィング部３０５に指示する。さらに、声質指定部１０４はその割合を時系列に沿って変化させる。
本実施の形態における音声モーフィング部３０５は、複数の音声合成部３０３から出力される合成音波形データ６１を取得して、その中間的性質を持つ中間的合成音波形データ１２を生成して出力する。
【００６２】

図１４は、本実施の形態における音声モーフィング部３０５の処理動作を説明するための説明図である。

本実施の形態における音声モーフィング部３０５は波形編集部３０５ａを備えている。

この波形編集部３０５ａは、声質指定部１０４により指定された少なくとも２つの合成音波形データ６１と割合とを特定し、それらの合成音波形データ６１から、その割合に応じた中間的合成音波形データ１２を生成する。

【００６３】

即ち、波形編集部３０５ａは、複数の合成音波形データ６１から、声質指定部１０４により指定された２つ以上の合成音波形データ６１を選択する。そして、波形編集部３０５ａは、声質指定部１０４により指定された割合に応じ、その選択した合成音波形データ６１のそれぞれに対して、例えば各音声の各サンプリング時点におけるピッチ周波数や振幅、各音声における各有声区間の継続時間長などを変形する。波形編集部３０５ａは、そのように変形された合成音波形データ６１を重ね合わせることで、中間的合成音波形データ１２を生成する。

【００６４】

スピーカ１０７は、このように生成された中間的合成音波形データ１２を波形編集部３０５ａから取得して、その中間的合成音波形データ１２に対応する合成音声を出力する。

このように、本実施の形態においても、実施の形態１又は２と同様、声質の自由度が広く良い音質の合成音声をテキスト１０から生成することができる。

【００６５】

（実施の形態４）

図１５は、本発明の実施の形態４に係る音声合成装置の構成を示す構成図である。

本実施の形態の音声合成装置は、出力する合成音声の声質に応じた顔画像を表示するものであって、実施の形態１に含まれる構成要素と、複数の顔画像に関する画像情報を蓄積する複数の画像ＤＢ４０１ａ～４０１ｚと、これらの画像ＤＢ４０１ａ～４０１ｚに蓄積される顔画像の情報を用いて画像モーフィング処理を行い、中間的顔画像データ１２ｐを出力する画像モーフィング部４０５と、画像モーフィング部４０５から中間的顔画像データ１２ｐを取得して、その中間的顔画像データ１２ｐに応じた顔画像を表示する表示部４０７とを備えている。

【００６６】

画像ＤＢ４０１ａ～４０１ｚのそれぞれが蓄積する画像情報の示す顔画像の表情は異なっている。例えば、怒っている声質の音声合成ＤＢ１０１ａに対応する画像ＤＢ４０１ａには、怒っている表情の顔画像に関する画像情報が蓄積されている。また、画像ＤＢ４０１ａ～４０１ｚに蓄積されている顔画像の画像情報には、顔画像の眉及び口の端や中央、目の中心点など、この顔画像の表す表情の印象をコントロールするための特徴点が付加されている。

【００６７】

画像モーフィング部４０５は、声質指定部１０４により指定された各合成音声パラメタ値列１０２のそれぞれの声質に対応付けされた画像ＤＢから画像情報を取得する。そして、画像モーフィング部４０５は、取得した画像情報を用いて、声質指定部１０４により指定された割合に応じた画像モーフィング処理を行う。

具体的に、画像モーフィング部４０５は、取得した一方の画像情報により示される顔画像の特徴点の位置が、声質指定部１０４により指定された割合だけ、取得した他方の画像情報により示される顔画像の特徴点の位置に変位するように、その一方の顔画像をワーピングし、これと同様に、その他方の顔画像の特徴点の位置を、声質指定部１０４により指定された割合だけ、その一方の顔画像の特徴点の位置に変位するように、その他方の顔画像をワーピングする。そして、画像モーフィング部４０５は、ワーピングされたそれぞれの顔画像を、声質指定部１０４により指定された割合に応じてクロスディゾルブすることで、中間的画像データ１２ｐを生成する。

【００６８】

これにより本実施の形態では、例えばエージェントの顔画像と合成音声の声質の印象を常に一致させることができる。即ち、本実施の形態の音声合成装置は、エージェントの平常声と怒り声の間の音声モーフィングを行って、少しだけ怒った声質の合成音声を生成するときには、音声モーフィングと同様の比率でエージェントの平常顔画像と怒り顔画像の間の画像モーフィングを行い、エージェントのその合成音声に適した少しだけ怒った顔画像を表示する。言い換えれば、感情を持つエージェントに対してユーザが感じる聴覚的印象と、視覚的印象を一致させることができ、エージェントの提示する情報の自然性を高めることができる。

【００６９】

図１６は、本実施の形態の音声合成装置の動作を説明するための説明図である。

例えば、ユーザが声質指定部１０４を操作することにより、図３に示すディスプレイ上の指定アイコン１０４ｉを、声質アイコン１０４Ａと声質アイコン１０４Ｚを結ぶ線分を４：６に分割する位置に配置すると、音声合成装置は、スピーカ１０７から出力される合成音声が１０％だけ声質Ａ寄りになるように、その４：６の割合に応じた音声モーフィング処理を声質Ａ及び声質Ｚの音声合成パラメタ値列１１を用いて行い、声質Ａ及び声質Ｂの中間的な声質ｘの合成音声を出力する。これと同時に、音声合成装置は、上記割合と同じ４：６の割合に応じた画像モーフィング処理を、声質Ａに対応付けられた顔画像Ｐ１と、声質Ｚに対応付けられた顔画像Ｐ２とを用いて行い、これらの画像の中間的な顔画像Ｐ３を生成して表示する。ここで、音声合成装置は、画像モーフィングするときには、上述のように、顔画像Ｐ１の眉や口の端などの特徴点の位置を、顔画像Ｐ２の眉や口の端などの特徴点の位置に向けて４０％の割合で変化するように、その顔画像Ｐ１をワーピングし、これと同様に、顔画像Ｐ２の特徴点の位置を、顔画像Ｐ１の特徴点の位置に向けて６０％の割合で変化するように、その顔画像Ｐ２をワーピングする。そして、画像モーフィング部４０５は、ワーピングされた顔画像Ｐ１に対して６０％の割合で、ワーピングされた顔画像Ｐ２に対して４０％の割合でクロスディゾルブし、その結果、顔画像Ｐ３を生成する。

【００７０】

このように、本実施の形態の音声合成装置は、スピーカ１０７から出力する合成音声の声質が「怒っている」ときには、「怒っている」様子の顔画像を表示部４０７に表示し、声質が「泣いている」ときには、「泣いている」様子の顔画像を表示部４０７に表示する。さらに、本実施形態の音声合成装置は、その声質が「怒っている」ものと「泣いている」ものとの中間的なものであるときには、「怒っている」顔画像と「泣いている」顔画像の中間的な顔画像を表示するとともに、その声質が「怒っている」ものから「泣いている」ものへと時間的に変化するときには、中間的な顔画像をその声質に一致させて時間的に変化させる。

【００７１】

なお、画像モーフィングは他にも様々な方法によって可能であるが、元となる画像の間の比率を指定することで目的の画像が指定できる方法であれば、どんなものを用いてもよい。

【産業上の利用可能性】

【００７２】

本発明は、声質の自由度が広く良い音質の合成音声をテキストデータから生成することができるという効果を有し、ユーザに対して感情を表す合成音声を出力する音声合成装置などに適用することができる。

【図面の簡単な説明】

【００７３】

【図１】本発明の実施の形態１に係る音声合成装置の構成を示す構成図である。

【図２】同上の音声合成部の動作を説明するための説明図である。

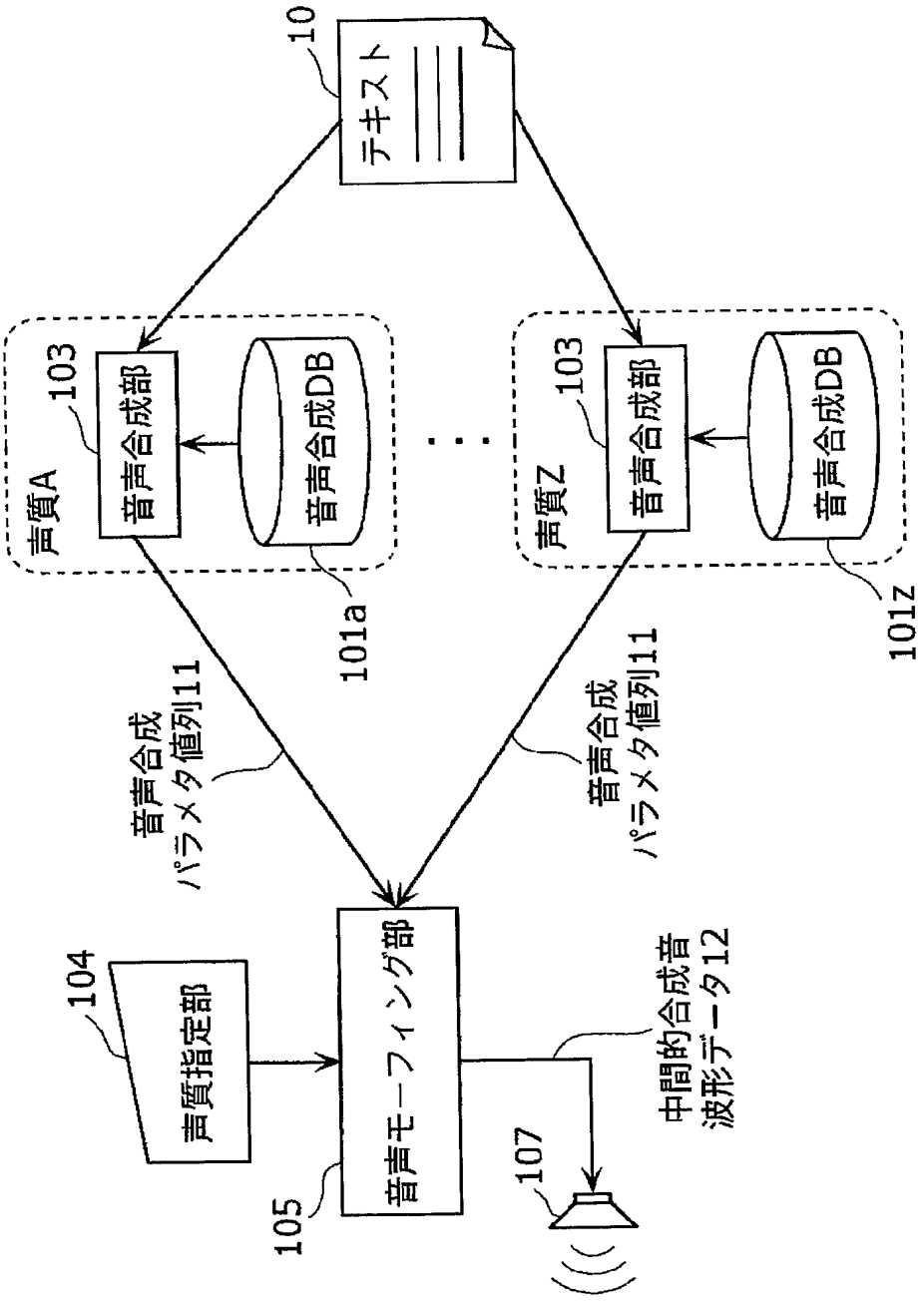
【図３】同上の声質指定部のディスプレイが表示する画面の一例を示す画面表示図である。

【図４】同上の声質指定部のディスプレイが表示する他の画面の一例を示す画面表示図である。
【図５】同上の音声モーフィング部の処理動作を説明するための説明図である。
【図６】同上の音声素片とＨＭＭ音素モデルの一例を示す例示図である。
【図７】同上の変形例に係る音声合成装置の構成を示す構成図である。
【図８】本発明の実施の形態２に係る音声合成装置の構成を示す構成図である。
【図９】同上の音声モーフィング部の処理動作を説明するための説明図である。
【図１０】同上の声質Ａ及び声質Ｚの合成音スペクトルと、それらに対応する短時間フーリエスペクトルとを示す図である。
【図１１】同上のスペクトルモーフィング部が両短時間フーリエスペクトルを周波数軸上で伸縮する様子を説明するための説明図である。
【図１２】同上のパワーが変換された２つの短時間フーリエスペクトルを重ね合わせる様子を説明するための説明図である。
【図１３】本発明の実施の形態３に係る音声合成装置の構成を示す構成図である。
【図１４】同上の音声モーフィング部の処理動作を説明するための説明図である。
【図１５】本発明の実施の形態４に係る音声合成装置の構成を示す構成図である。
【図１６】同上の音声合成装置の動作を説明するための説明図である。

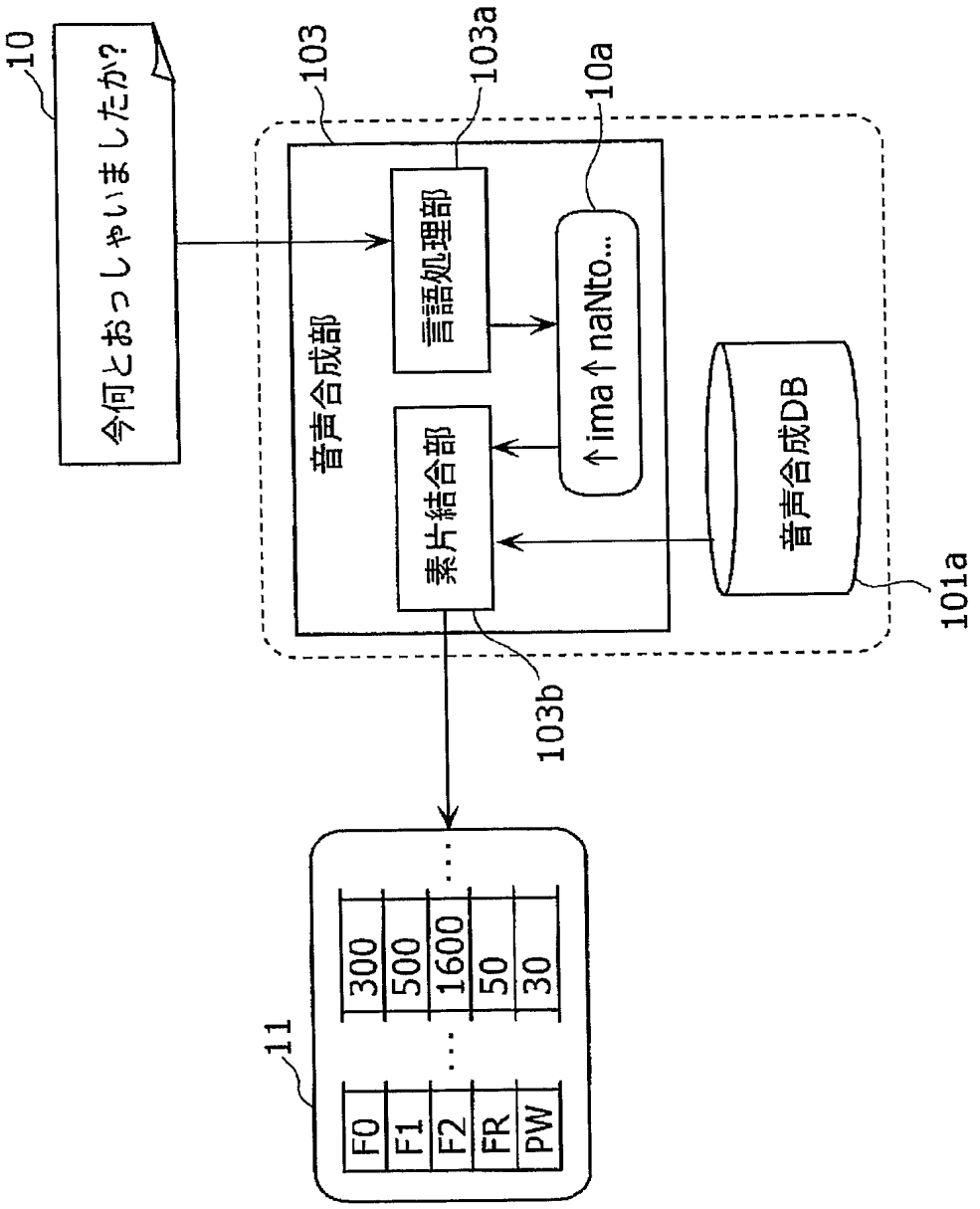
【符号の説明】
【００７４】
１０　テキスト
１０ａ　音素情報
１１　音声合成パラメタ値列
１２　中間的合成音波形データ
１２ｐ　中間的画像データ
１３　中間的音声合成パラメタ値列
３０　音声素片
３１　音素モデル
３２　最尤パスの形状
４１　合成音スペクトル
４２　中間的合成音スペクトル
５０　フォルマント形状
５０ａ，５０ｂ　周波数
５１　フーリエスペクトル分析窓
６１　合成音波形データ
１０１ａ～１０１ｚ　音声合成ＤＢ
１０３　音声合成部
１０３ａ　言語処理部
１０３ｂ　素片結合部
１０４　声質指定部
１０４Ａ，１０４Ｂ，１０４Ｚ　声質アイコン
１０４ｉ　指定アイコン
１０５　音声モーフィング部
１０５ａ　パラメタ中間値計算部
１０５ｂ　波形生成部
１０６　中間的合成音波形データ
１０７　スピーカ
２０３　音声合成部
２０１ａ～２０１ｚ　音声合成ＤＢ
２０５　音声モーフィング部
２０５ａ　スペクトルモーフィング部

２０５ｂ　波形生成部
３０３　音声合成部
３０１ａ～３０１ｚ　音声合成ＤＢ
３０５　音声モーフィング部
３０５ａ　波形編集部
４０１ａ～４０１ｚ　画像ＤＢ
４０５　画像モーフィング部
４０７　表示部
Ｐ１～Ｐ３　顔画像

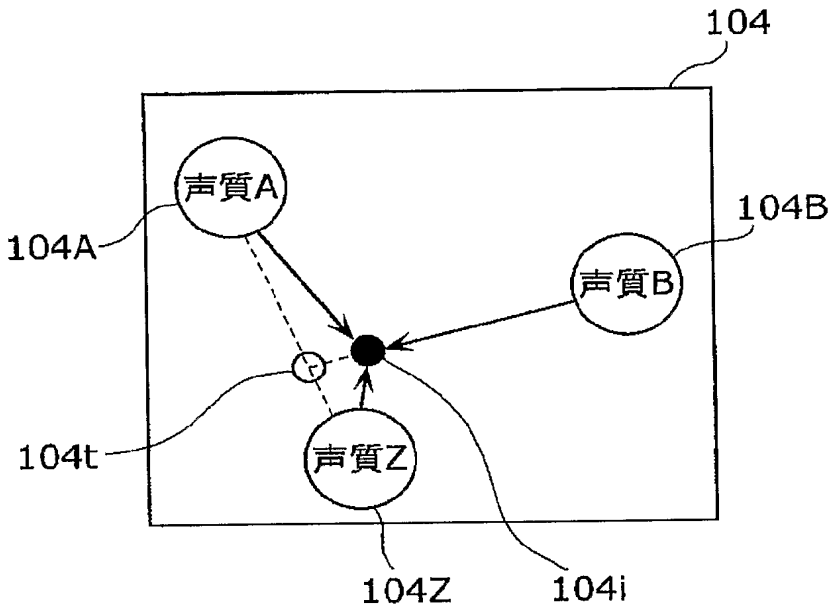
【書類名】図面
【図１】


10
テキスト
声質A
103
音声合成部
音声合成DB
101a
声質Z
103
音声合成部
音声合成DB
101z
音声合成
パラメタ値列11
音声合成
パラメタ値列11
104
声質指定部
音声モーフィング部
105
107
中間的合成音
波形データ12

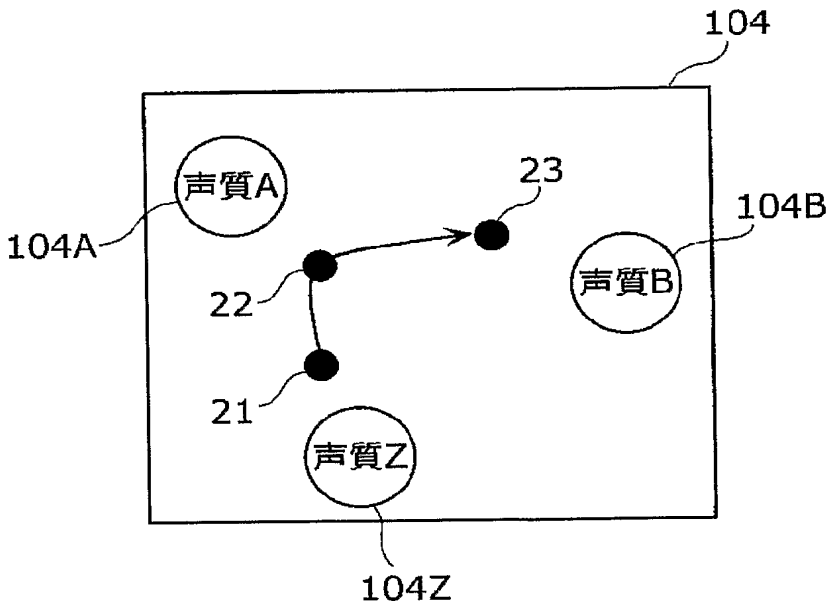
【図２】


10
今何とおっしゃいましたか?
103
103a
10a
音声合成部
言語処理部
素片結合部
↑ima↑naNto...
音声合成DB
101a
103b
11
300
500
1600
50
30
F0
F1
F2
FR
PW

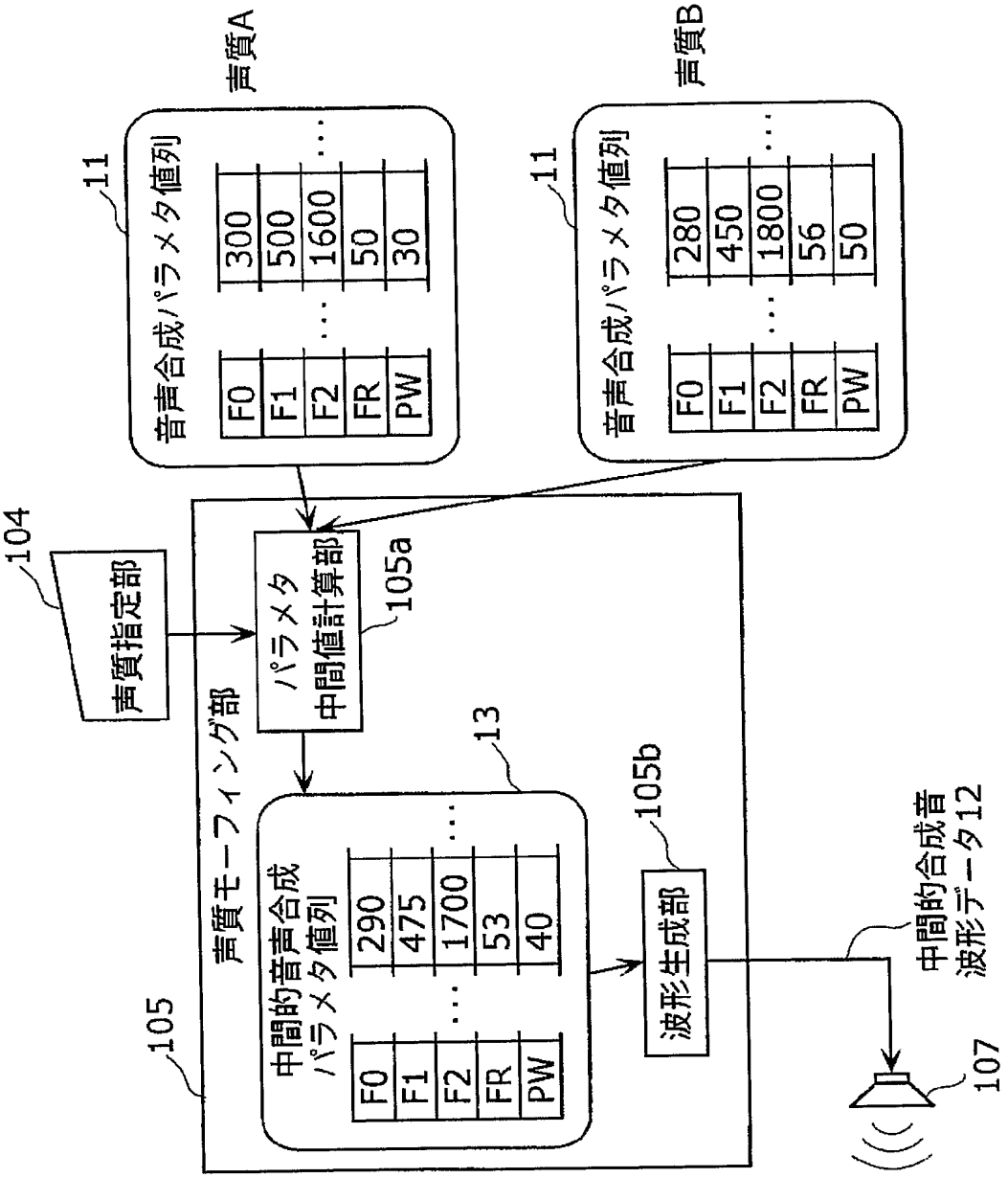
【図３】


104
声質A
104A
104B
声質B
104t
声質Z
104Z
104i

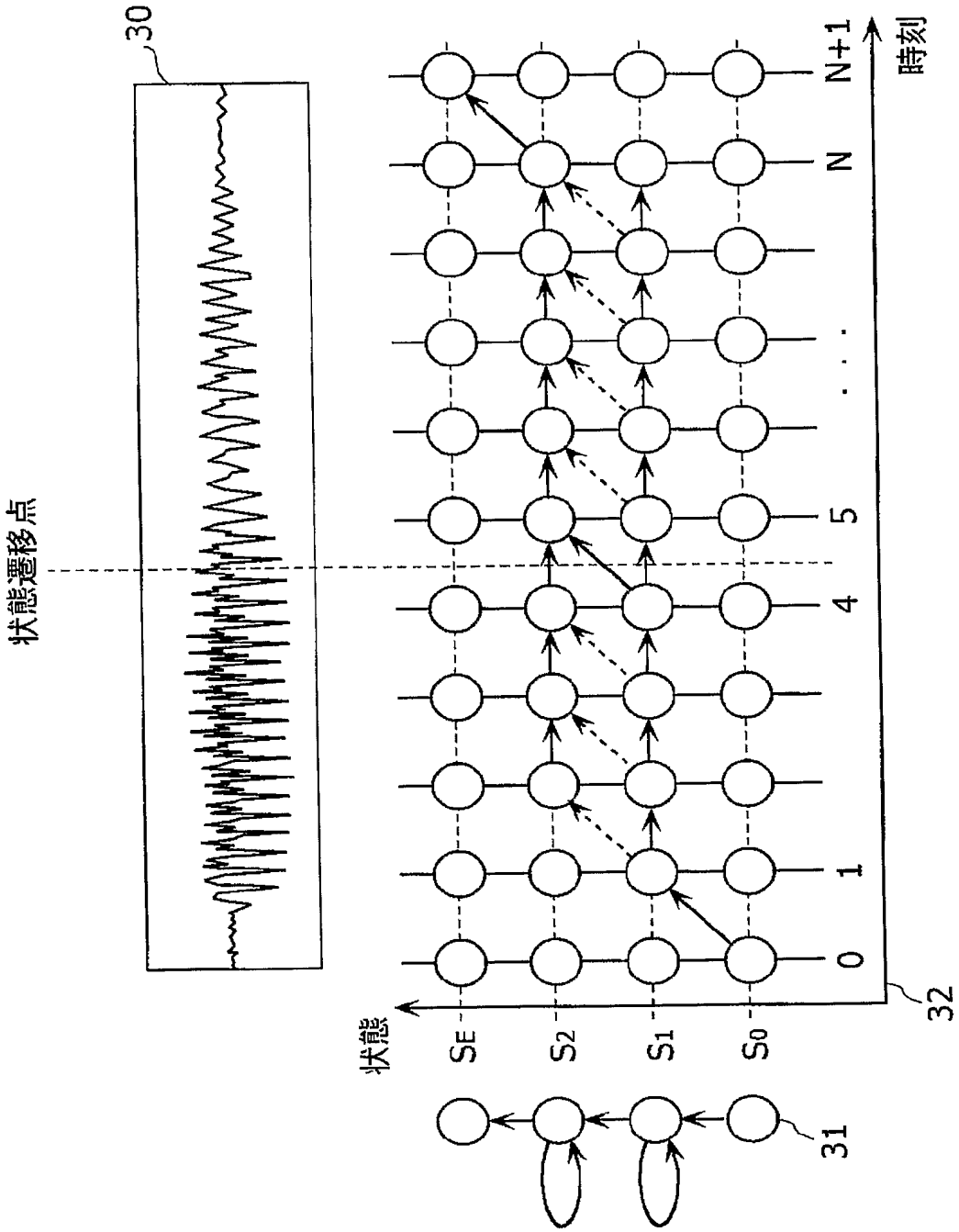
【図４】


104
声質A
104A
23
104B
声質B
22
21
声質Z
104Z

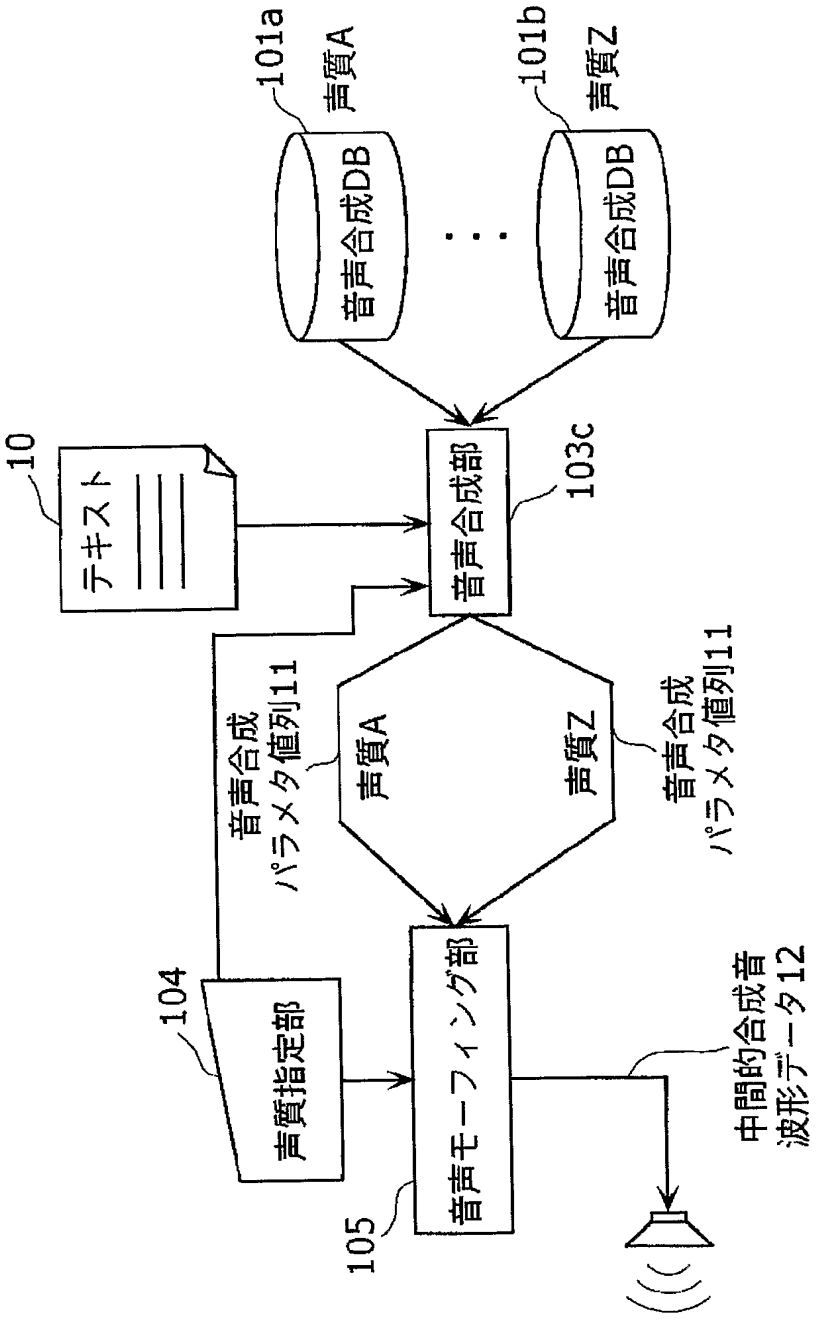
【図５】


声質A
11
音声合成パラメタ値列
F0
F1
F2
FR
PW
300
500
1600
50
30
声質B
11
音声合成パラメタ値列
F0
F1
F2
FR
PW
280
450
1800
56
50
104
声質指定部
パラメタ
中間値計算部
105a
声質モーフィング部
13
中間的音声合成
パラメタ値列
F0
F1
F2
FR
PW
290
475
1700
53
40
105b
波形生成部
105
中間的合成音
波形データ12
107

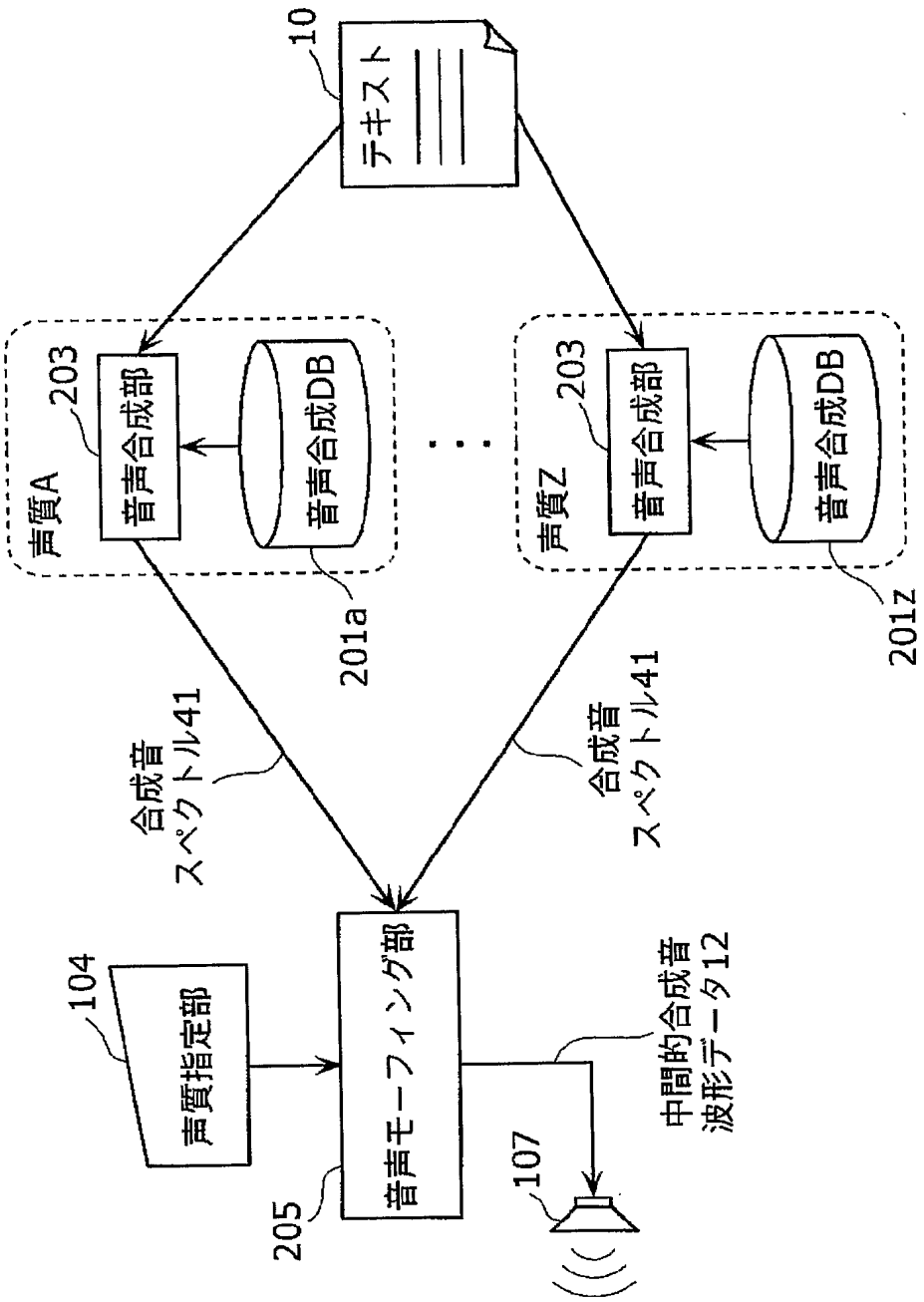
【図６】



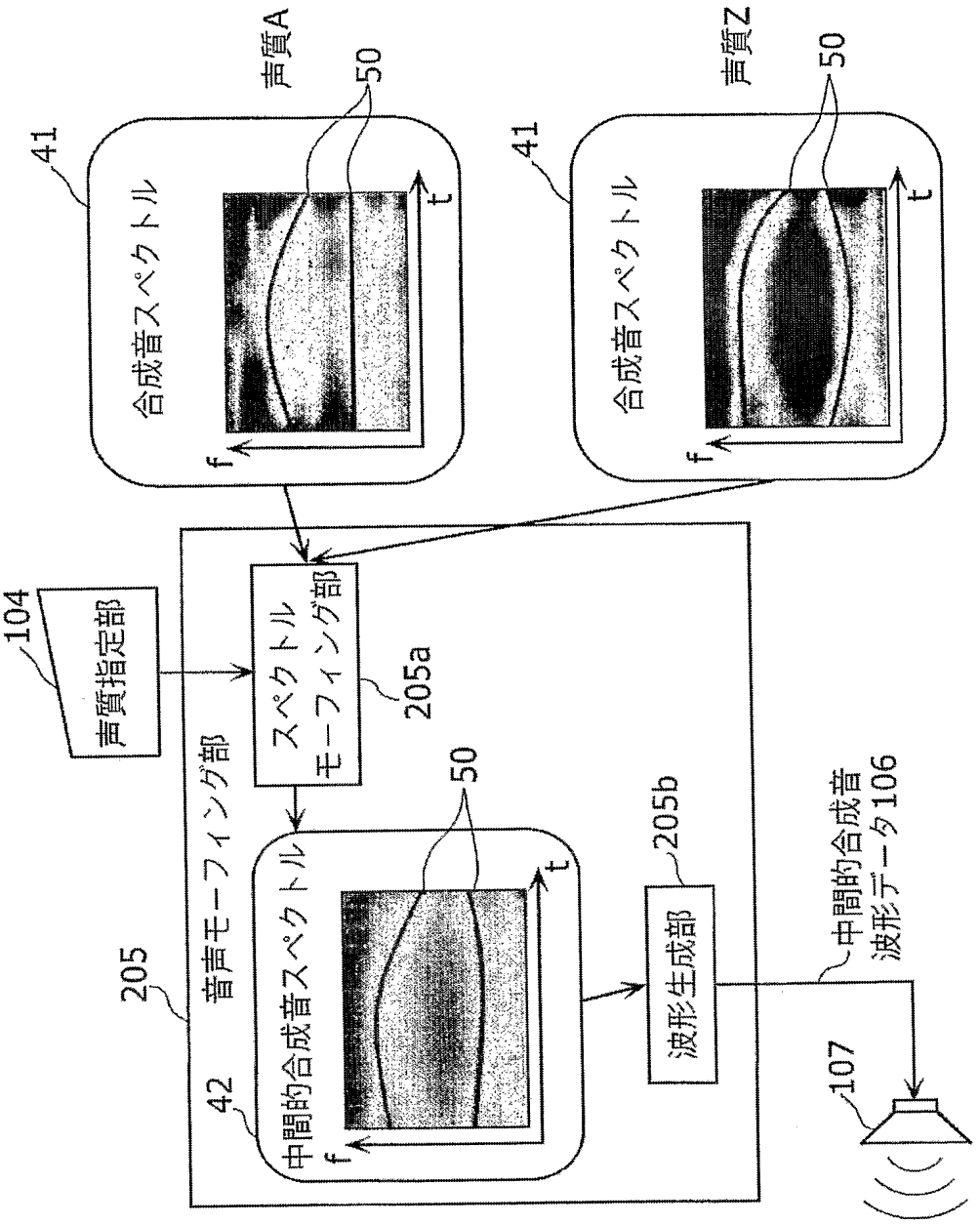
【図７】


101a
声質A
音声合成DB
101b
声質Z
音声合成DB
10
テキスト
音声合成部
103c
音声合成
パラメタ値列11
声質A
声質Z
音声合成
パラメタ値列11
104
声質指定部
音声モーフィング部
105
中間的合成音
波形データ12

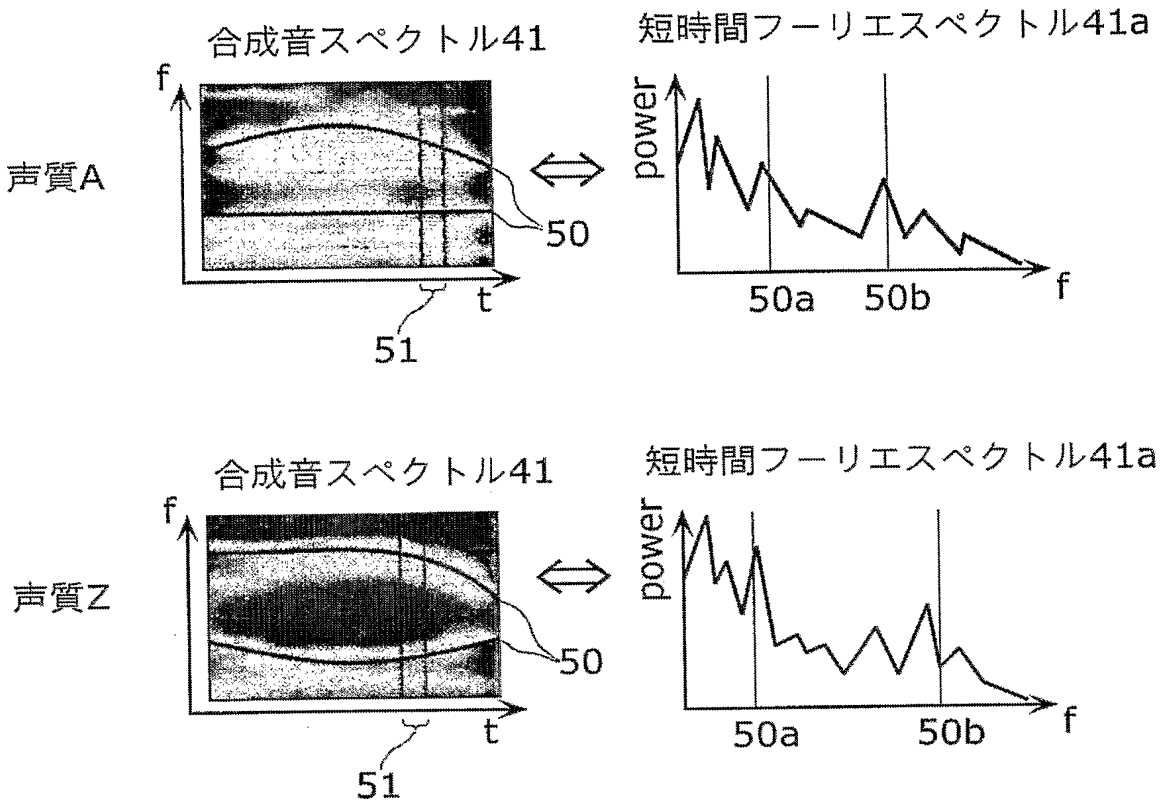
【図８】


10
テキスト
声質A
203
音声合成部
音声合成DB
201a
声質Z
203
音声合成部
音声合成DB
201z
合成音
スペクトル41
合成音
スペクトル41
104
声質指定部
音声モーフィング部
205
107
中間的合成音
波形データ12

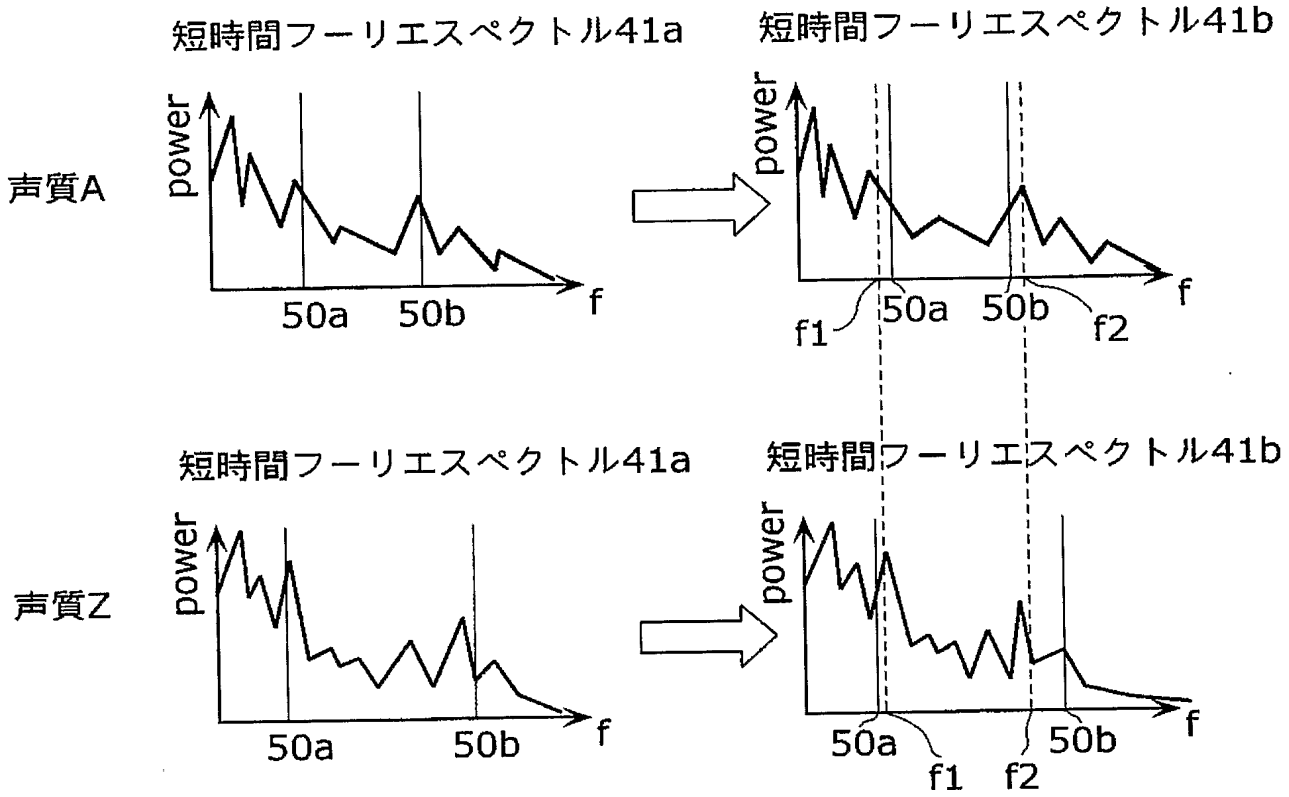
【図９】


声質A
50
41
合成音スペクトル
t
f
声質Z
50
41
合成音スペクトル
t
f
104
声質指定部
スペクトル
モーフィング部
205a
音声モーフィング部
205
50
中間的合成音スペクトル
t
f
42
205b
波形生成部
中間的合成音
波形データ106
107

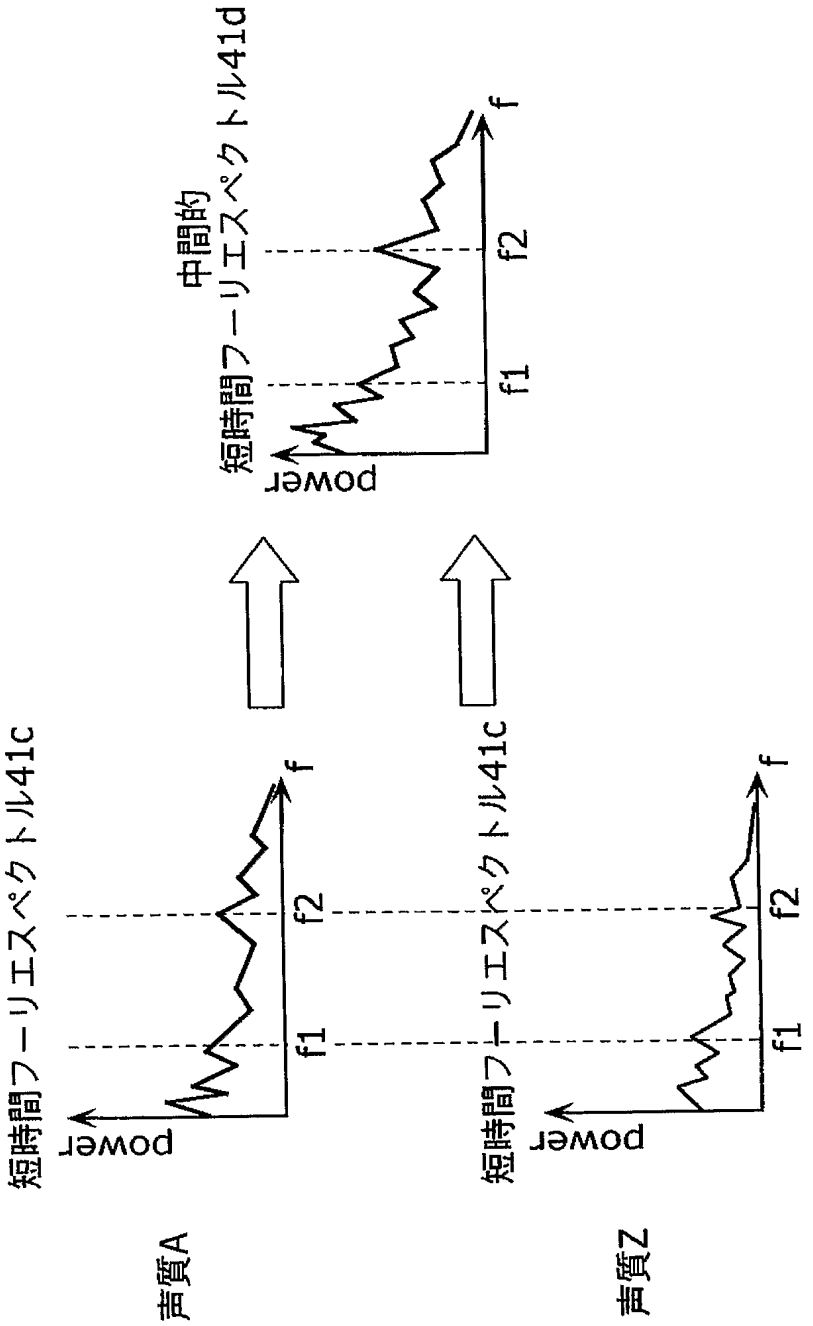
【図１０】


合成音スペクトル41
短時間フーリエスペクトル41a
声質A
f
t
50
51
power
50a
50b
f
合成音スペクトル41
短時間フーリエスペクトル41a
声質Z
f
t
50
51
power
50a
50b
f

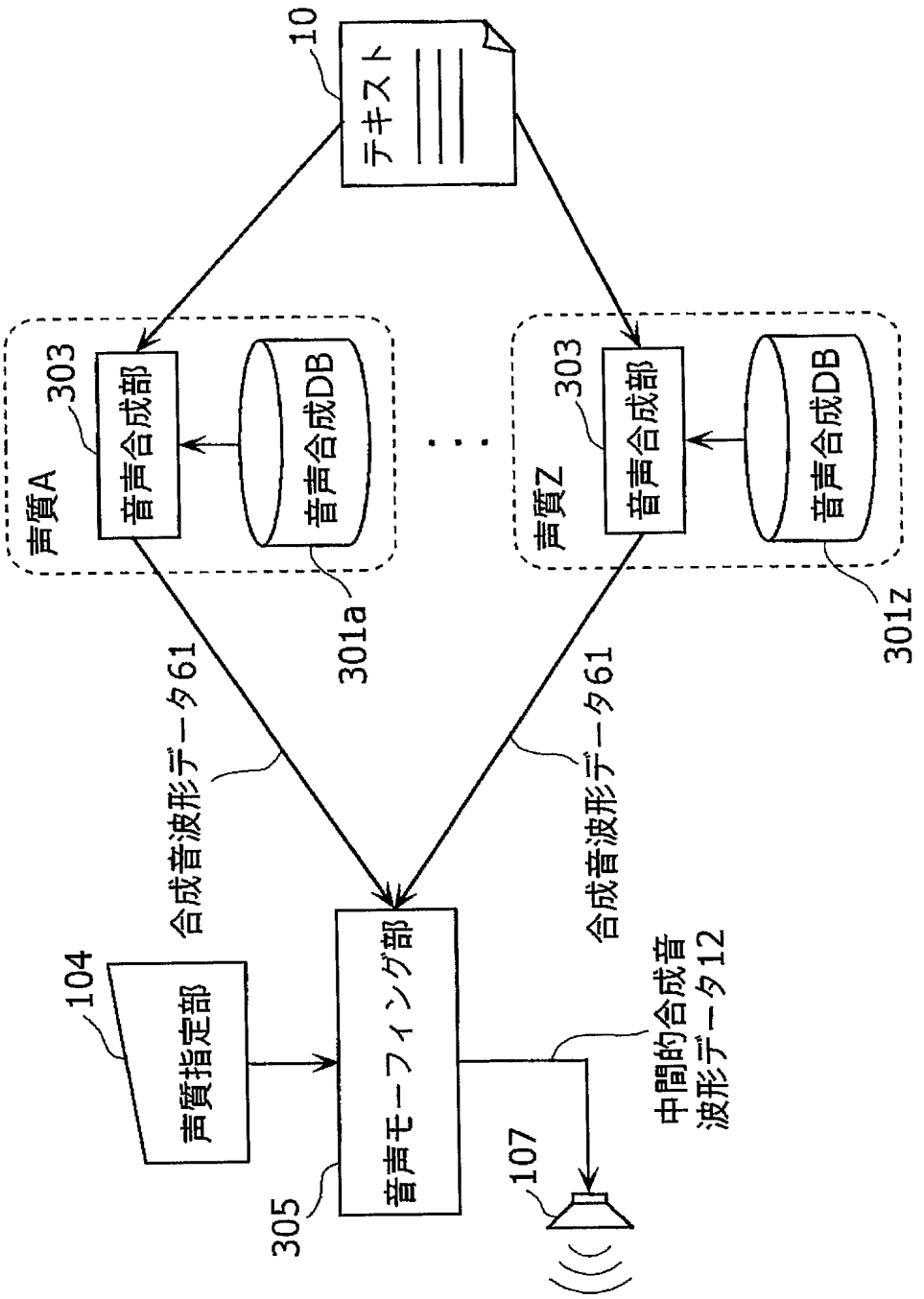
【図１１】


短時間フーリエスペクトル41a
短時間フーリエスペクトル41b
声質A
power
50a
50b
f
f1
f2
短時間フーリエスペクトル41a
短時間フーリエスペクトル41b
声質Z
power
50a
50b
f
f1
f2

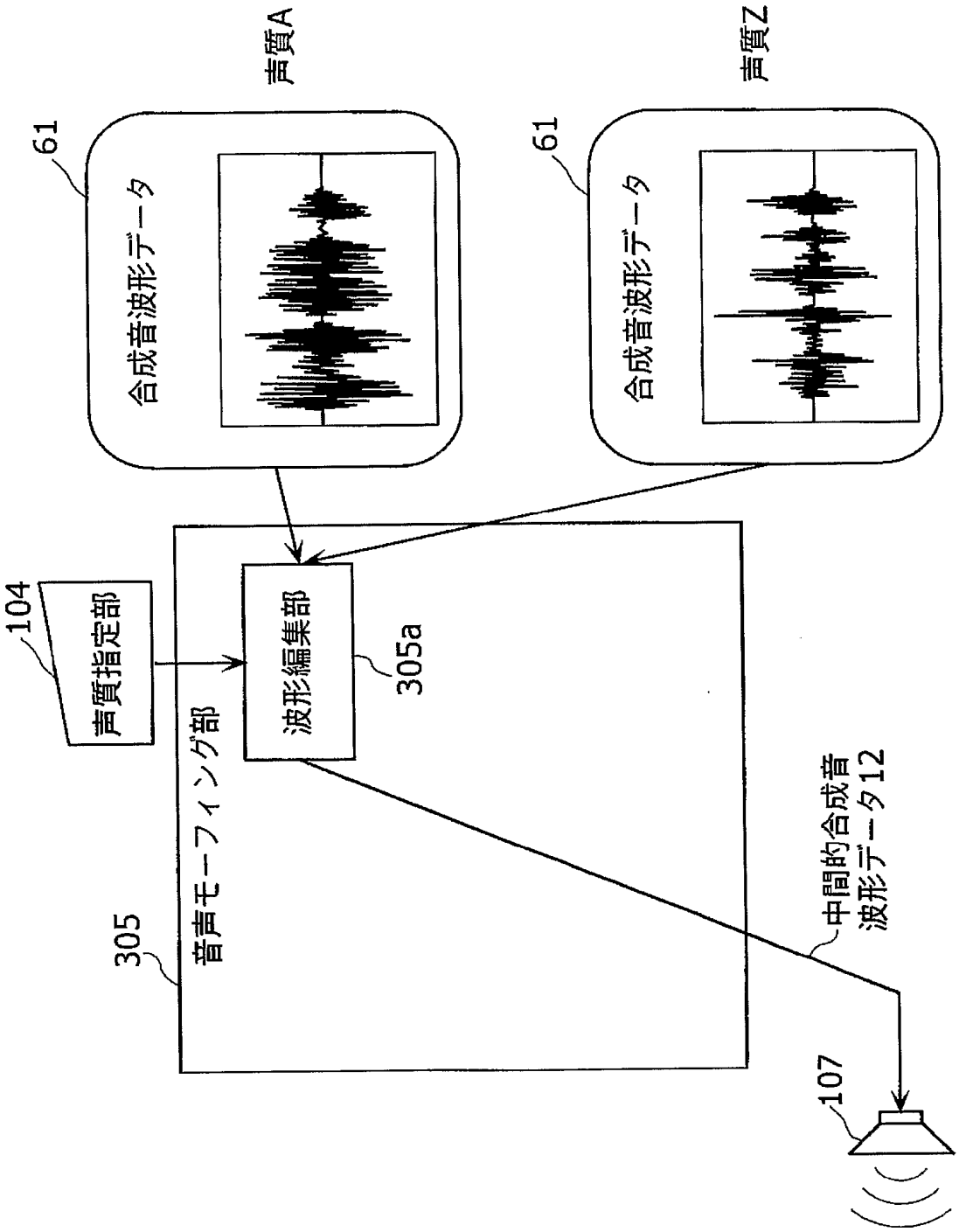
【図１２】


短時間フーリエスペクトル41c
power
f
f1
f2
声質A
短時間フーリエスペクトル41c
power
f
f1
f2
声質Z
中間的
短時間フーリエスペクトル41d
power
f
f1
f2

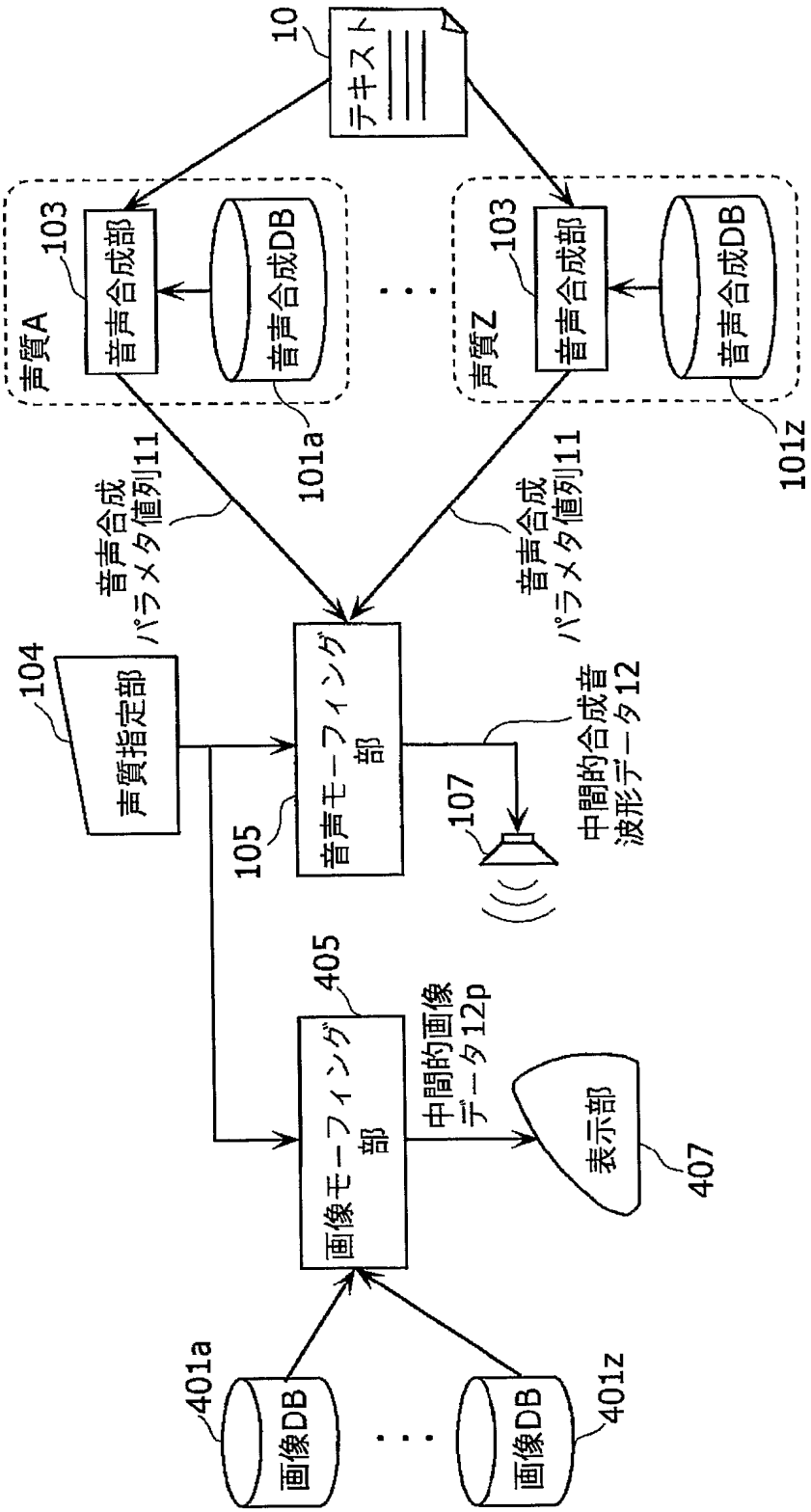
【図１３】


10
テキスト
声質A
303
音声合成部
音声合成DB
301a
声質Z
303
音声合成部
音声合成DB
301z
合成音波形データ61
合成音波形データ61
104
声質指定部
音声モーフィング部
305
107
中間的合成音
波形データ12

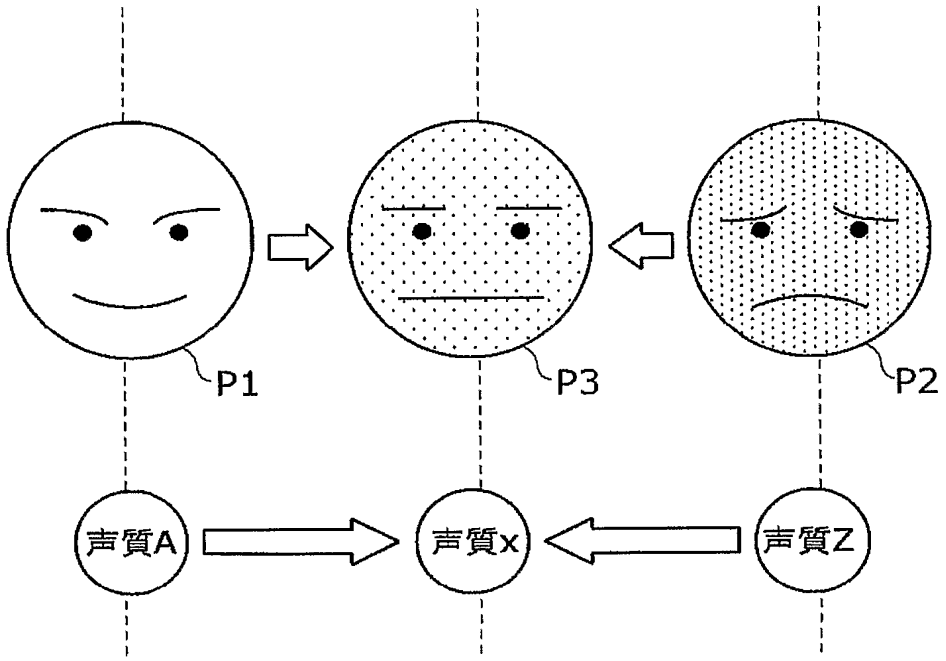
【図１４】


声質A
声質Z
61
61
合成音波形データ
合成音波形データ
104
声質指定部
波形編集部
305a
305
音声モーフィング部
中間的合成音
波形データ12
107

【図１５】


10
テキスト
声質A
103
音声合成部
音声合成DB
101a
声質Z
103
音声合成部
音声合成DB
101z
音声合成
パラメタ値列11
音声合成
パラメタ値列11
104
声質指定部
105
音声モーフィング
部
107
中間的合成音
波形データ12
405
画像モーフィング
部
中間的画像
データ12p
表示部
407
401a
画像DB
401z
画像DB

【図１６】



【書類名】要約書
【要約】
【課題】　声質の自由度が広く良い音質の合成音声をテキストデータから生成する音声合成装置を提供することを目的とする。
【解決手段】　音声合成装置は、音声合成ＤＢ１０１ａ，１０１ｚと、テキスト１０を取得するとともに、音声合成ＤＢ１０１ａから、テキスト１０に含まれる文字に対応した声質Ａの音声合成パラメタ値列１１を生成する音声合成部１０３と、音声合成ＤＢ１０１ｚから、テキスト１０に含まれる文字に対応した声質Ｚの音声合成パラメタ値列１１を生成する音声合成部１０３と、声質Ａ及び声質Ｚの音声合成パラメタ値列１１から、テキスト１０に含まれる文字に対応した、声質Ａ及び声質Ｚの中間的な声質の合成音声を示す中間的音声合成パラメタ値列１３を生成する音声モーフィング部１０５と、生成された中間的音声合成パラメタ値列１３をその合成音声に変換して出力するスピーカ１０７とを備える。
【選択図】　図１

認定・付加情報

特許出願の番号	特願２００４－０１８７１５
受付番号	５０４００１３３２７７
書類名	特許願
担当官	第八担当上席　　００９７
作成日	平成１６年　１月２８日

＜認定情報・付加情報＞

【提出日】　平成16年 1月27日

特願２００４－０１８７１５

出願人履歴情報

識別番号　［０００００５８２１］

１．変更年月日　１９９０年　８月２８日
［変更理由］　新規登録
住　所　大阪府門真市大字門真１００６番地
氏　名　松下電器産業株式会社